Exploring the Phylodynamics, Genetic Reassortment and RNA Secondary Structure formation patterns of Orthomyxoviruses by Comparative Sequence Analysis

By Fredrick Nzabanyi Nindo
(BSc, MSc)

Supervisor
Assoc Prof Darren Martin

Thesis presented for the degree of

DOCTOR OF PHILOSOPHY

In the Department of Integrative Biomedical Sciences

Faculty of Health Sciences,

UNIVERSITY OF CAPE TOWN

November 2019

ACKNOWLEDGEMENTS

My first acknowledgements goes to my supervisor Associate Professor Darren Patrick Martin for giving me an opportunity to pursue PhD under guidance and encouragement throughout the research period without which this work would not have come to completion. I particularly wish to thank him for giving me the freedom to explore research agenda of my choice but with his mentorship. I wish also to acknowledge the Darren Lab's members especially Dr Brejnev Muhire for his help with debugging of some scripts when the going got tough and not forgetting Dr Gordon Harkins of SANBI, UWC, for his useful discussions on the emerging approaches and techniques in phylodynamics in the era of sequence data deluge.

My sincere thanks also goes to the members of CBIO group who are many to name individually here as computational biology research family, they played a role in the realization of this work. Further, acknowldgements also goes to Ayton and Gerrit (the senior systemsdeveloper and Bioinformatics engineer) who came to my rescue by offering me computation space on the CBIO cluster when I needed it most to bring to completion my analyses.

I thank God for the wonderful parents and members of my extended family in Kenya who have always believed in me, for their encouragement to always aim high. I also thank God for wonderful friends and the Cape Town community that I have had wonderful interactions and who played a role to make my stay hospitable during the course of my studies. May you remain blessed.

Last but not least, I wish to thank the Carnegie Corporation of New York for their generous funding support for the prestigious PhD scholarship for research in Infectious Diseases under the Next Generation of Academics in Africa (NGAA) programme for tenable at the Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town and the Polimyelitis Research Foundation (PRF) of South Africa for their PhD Bursary for research in Medical Virology in South Africa.

DEDICATION

I dedicate this dissertation to my lovely wife, **Eva** and courageous sons; **Keith** and **Danny** in for their love, resilience and and unlimited support during my entire study period.

DECLARATION

I, **Fredrick Nzabanyi Nindo,** hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being,or is to be submitted for another degree in this or any other university. I authorize the University to reproduce for the purpose of research either whole or any portion of the contents in any manner whatsoever.

Signature**:…** Signed by candidate Date**:** 12.02.2020

ABSTRACT

RNA viruses are among the most virulent microorganisms that threaten the health of humans and livestock. Among the most socio-economically important of the known RNA viruses are those found in the family Orthomyxovirus. In this era of rapid low cost genome sequencing and advancements in computational biology techniques, many previously difficult research questions relating to the molecular epidemiology and evolutionary dynamics of these viruses can now be answered with ease. Using sequence data together with associated meta-data, in chapter two of this dissertation I tested the hypothesis that the Influenza A/H1N1 2009 pandemic virus was introduced multiple times into Africa, and subsequently dispersed heterogeneously across the continent. I further tested to what degree factors such as road distances and air travel distances impacted the observed pattern of spread of this virus in Africa using a generalised linear model based approach.

In chapter three, I set out to test two hypotheses: (1) that there is no difference in the frequency of reassortments among the segments that constitute influenza virus genomes; and (2) that there is epochal temporal reassortment among influenza viruses and that all geographical regions are equally likely sources of epidemiologically important influenza virus reassortant lineages

In chapter four of this thesis, I explored the formation of RNA secondary structures within the genomes of orthomyxoviruses belonging to five genera: Influenza A, B and C, Infectious Salmon Anaemia Virus and Thogotovirus using *in silico* RNA folding predictions and additional molecular evolution and phylogenetic tests to show that structured regions may be biologically functional. The presence of some conserved structures across the five genera is likely a reflection of the biological importance of these structures, warranting further investigation regarding their role in the evolution and possible development of antiviral resistance.

The studies herein demonstrate that pathogen genomics-based analytical approaches are useful both for understanding the mechanisms that drive the evolution and spread of rapidly evolving viral pathogens such as orthomyxoviruses, and for illuminating how these approaches could be leveraged to improve the management of these pathogens.

CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

LIST OF APPENDICES

LIST OF ACRONYMS

ANISE -African Network for Influenza Surveillance and Epidemiology

BEAST -Bayesian Evolutionary Analysis by Sampling Trees

BEAUti – Bayesian Evolutionary analysis Utility

BSSVS -Bayesian Stochastic Search Variable Selection

DNA – Deoxyribonucleic Acid

GDP – Gross Domestic Product

GISAID – Global Initiative for Shairing All Influenza Data

HA – Haemagglutinin

HCSS – High Confidence Secondary Structure

HE – Haemagglutinin Esterase

HEF – Haemagglutinin Esterase Fusion

HPR – Highly Polymorphic region

Isavirus -Infectious Salmon Anaemia Virus

MERS -Middle East Respiratory Syndrome

NA - Neuraminidase

NASP- Nucleic Acid Structure Predictor

NCBI – National Center for Biotechnology Information

NP - Nucleoprotein

RNA – Ribonucleic Acid

RNP -Ribonucleoproteins

SARS -Severe Acute Respiratory Syndrome

SPREAD3 - Spatial Phylogenetic Reconstruction of Evolutionary Dynamics using Data-Driven Documents (D3)

THOV – Thogoto virus

tMRCA – time to Most Recent Common Ancestor

WHO – World Health Organization

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Viral epidemics continue to pose serious human health threats especially with the increasing frequency of these epidemics (Marra, 2003; Rao, 2009; Christman *et al.*, 2011; T. T.-Y. Lam *et al.*, 2013; Saéz *et al.*, 2014; Faria *et al.*, 2016). A range of factors are potentially responsible for the emergence of these epidemics including. surging global human populations, climate change, increased volumes of international trade and increasing artes of long-distance human travel (Hay *et al.*, 2013; Klepac *et al.*, 2014; Rajao *et al.*, 2014; Joseph *et al.*, 2017). In recent years, emerging and re-emerging viruses have been shown to be the aetiologies of regional and global epidemics (Zappa *et al.*, 2009; Devaux, 2012; Bloom, Black and Rappuoli, 2017).

Several high-profile viral epidemics have occurred over the past two decades including Zika, Ebola, pandemic Influenza, MERS and SARS (Jombart *et al.*, 2009; Assiri *et al.*, 2013; T. T.-Y. Lam *et al.*, 2013; Gire *et al.*, 2014; Mayer, Tesh and Vasilakis, 2017) . These outbreaks started from single foci and rapidly grew to a global scale. The rapid global spread of many of these pathogens has made it clear that the world is insufficiently prepared to contain large-scale viral epidemics. Most African countries are especially disadvantaged when it comes to countering these epidemics because of their limited budgetary allocations to disease survaillance, lack of technical capacity, inadequate personnel and a lack of infrastructure (Gilsdorf, Morgan and Leitmeyer, 2012; Heraud *et al.*, 2012; Katz *et al.*, 2012; Saéz *et al.*, 2014; Venter, 2018) . As a consequence, only a tiny proportion of the available data on viral epidemic mitigation has been generated within an African context and relates directly to Africa (Bao *et al.*, 2008; Elbe and Buckland-Merrett, 2017). The studies conducted here focus on a small family of segmented negative stranded RNA viruses known as orthomyxoviruses (Lamb, Krug and Knipe, 2001). This family of viruses includes some of the most important human pathogens i.e. Influenza type A, B and C viruses as well as Infectious salmon anaemia virus (Isavirus) and the arbovirus Thogoto Virus (Falk *et al.*, 1997; M B Leahy *et al.*, 1997; Contreras-Gutiérrez *et al.*, 2017; Lefkowitz *et al.*, 2017).

1.1.1 Overall aims and Objectives

Overall aim

Explore the spatiotemporal and evolutionary dynamics of Orthomyxoviruses and evaluate the impact of these processes on their epidemiology.

Objectives

1. **Spatiotemporal transmission dynamics of the 2009 Influenza A/H1N1 pandemic virus in Africa**

To use publically available sequence data from dedicated Influenza sequence databases in combination with demographic, ecological and economic information of the sampling locations, to unravel the introduction to, and dissemination patterns across, Africa of the 2009 Influenza A/H1N1 pandemic virus.

Additionally, a number of demographic, ecological, economic and viral genetic factors were tested to describe the most critical contributors/predictors to the dispersal pattern of this virus in Africa. The results of these studies suggested that there were multiple introductions of these virus to Africa and that geographical distance and location latitude were the factors contributing most to the observed dispersal pattern during the height of the outbreak.

2. **Re-assortment analysis of Influenza A, B and C viruses**

Analysis of Influenza type A, B, and C virus sequence data spanning several decades (1927-2014) available in the public databases aimed at understanding the patterns of re-assortment and evolutionary dynamics in these three Influenza virus species. Time-scaled phylogeographic analyses were carried to determine when and where the reassortant Influenza viruses arose.

3. **Impact of RNA folding on the evolution of viruses with segmented genomes (orthomyxoviruses).**

The detection of possible biologically functional RNA folding structures within the genomes of Orthomyxoviruses and assessing the impacts of these structures on the mutational patterns of these viruses.

1.1.2 THESIS STRUCTURE

As with all RNA viruses, orthomyxoviruses have high mutation rates given the lack of proof reading during replication by RNA dependent RNA polymerases (Kühnert *et al.*, 2014). Thus the evolution of these viruses occurs on the same timescale as ecological and demographic processes: a factor making these viruses amenable to phylodynamic analyses. Phylodynamics has been used to precisely describe the evolution and transmission dynamics of several viruses . In chapter two, I attempt to use this framework, to elucidate the introduction, establishment and spread of the 2009 Influenza A/H1N1 pandemic virus in Africa. I further, extend this study to explore genetic, environmental, demographic, and economic factors that have most profoundly contributed to the observed transmission patterns of this virus on the African continent.

In chapter three, I explored patterns of genome-component re-assortment and attempt to reconstruct the emergence (where and when) of re-assortant Influenza virus lineages using concatenated whole genome sequences of Influenza A, B and C viruses isolated over the last 70 years (1927-2013). Orthomyxoviruses have segmented genomes that are packaged compactly to form a virion. In circumstances where two or more segmented virus strains infect a cell simultaneously; there is a chance of segment exchange between the distinct strains resulting in novel strains with mixed segment ancestry; a process referred to as reassortment. Reassortment contributes to the generation of considerable diversity in orthomyxoviruses and is most prominent in Influenza A viruses given the high degree of diversity and borad host range of this species.

In chapter four, I explore the computational prediction of RNA structure within homologous segments of Influenza A, B and C, Thogotovirus and Isavirus. I further perform further analyses that attempt to provide additional layers of evidence that paired sites within the genomes of these viruses are probably evolutionarily preserved and hence may be biologically functional. The multipartite nature of the orthomyxoviruses, poses a great challenge when it comes to packaging their genomes into single enveloped virions. This challenge is thought to be overcome by some regions of the segments forming structural modifications and conformations via intra-molecular base-pairing (i.e. RNA secondary structure formation).

In chapter five, I provide a summary of the inferences made from the various analyses performed and suggest additional studies that are warranted to counter the spread of infectious diseases during outbreaks, track epidemiological important re-assortant Influenza virus strains, and determine the mechanisms/strategies that could be employed to formulate gene therapeutic techniques targeting the conserved folding patterns within orthomyxoviruses.

1.2 LITERATURE REVIEW

1.2.0 OVERVIEW OF ORTHOMYXOVIRUS TAXONOMY

Orthomyxoviruses are a higher order classification of a family of single stranded segmented negative sense RNA viruses (Lamb, Krug and Knipe, 2001; Fields, Knipe and Howley, 2007). Their name derives from the Greek words '*orthos*' meaning "straight" and '*myxa*' meaning "mucus". The initial genera in this family of viruses included Influenza A virus, Influenza B virus, and Influenza C virus, but subsequently Infectious Salmon Anemia Virus (Isavirus/Isavirus) and Thogoto Virus (THOV) were respectively added to the family in the 1980's and early 1990's (Webster et al. 1992; Leahy et al. 1997; Mjaaland et al. 1997; Kibenge et al. 2004). Continued virus surveillance efforts have seen the further discovery of additional members of the orthomyxovirus family including the genus Quaranjavirus which has three species: Quaranfil, Johnston Atoll and Lake Chad viruses (Presti *et al.*, 2009) (Figure 1.2.0). Another addition to the family has been Influenza type D virus as new genus that is most closely related to Influenza C (Hause *et al.*, 2014). Although Influenza D viruses were initially isolated from swine, they have most commonly been found infecting cattle and are now frequently referred to as cattle Influenza (Murakami et al. 2016; Foni et al. 2017).

Orthomyxoviruses are currently classified into seven genera: Alphainfluenzavirus (e.g. Influenza A virus) , Betainfluenzavirus (e.g. Influenza B virus), Deltainfluenzavirus (e.g. Influenza D virus), Gammainfluenzavirus (e.g. Influenza C virus), Isavirus (e.e. Infectious Salmon Anemia Virus), Quaranjavirus (e.g. Quaranfil virus) and Thogotovirus (e.g. Thogoto virus;King *et al.*, 2018).

The Influenza viruses are genetically typed using the antigenicity of their nucleoprotein (NP) and matrix proteins (M1/M2) (Kobasa and Kawaoka, 2005). Influenza A virus is further subtyped based on the antigenicity of their surface proteins, Haemagglutinin (HA) and Neuraminidase (NA) (Kobasa and Kawaoka, 2005).

Influenza A and B viruses are characterized by eight genomic segments whereas Influenza C and D virus genomes have seven segments (Hause *et al.*, 2014; Matsuzaki *et al.*, 2016). Isavirus and Thogoto virus genomes span eight and six segments respectively (Falk *et al.*, 1997; M B Leahy *et al.*, 1997; Mjaaland *et al.*, 1997). Influenza A virus and Influenza B viruses have two surface proteins (HA and NA), Isavirus has two (HA and fusion protein; F), Influenza C and D viruses have one surface protein (haemagglutinin esterase; HEF), and Thogoto virus has one (glycoprotein) (Hilleman, 2002). Like all RNA viruses, mutational rates in orthomyxoviruses

are high (Pompei, Loreto and Tria, 2012). In addition to being generated through mutation, genetic diversity in orthomyxoviruses is also generated through reassortment (Vijaykrishna, Mukerji and Smith, 2015). An importnat impact of reassortment is that it is associated with host species changes such as those responsible for pandemics in humans (Assiri *et al.*, 2013; Saéz *et al.*, 2014; Lee *et al.*, 2015).

Some segments code for a single gene/protein whereas others undergo alternative splicing to generate more than one gene/protein (Ojosnegros *et al.*, 2011; Zemora *et al.*, 2016). The packaging of several segments into a single enveloped virion is mechanistically challenging and is achieved in part by the RNA molecules of each segment folding into more compact secondary and tertiary structures (A P Gultyaev *et al.*, 2014; Gerber *et al.*, 2014).

Replication of orthomyxoviruses is a function dedicated to the polymerase complex that is encoded on three segments (segments 1,2, and 3), whereas host cell attachment is facilitated by the surface glycoproteins coded for by segment 4 and 6 in flu A and B, the HEF segment in flu C, the HA and F segments in Isavirus and segment GP in THOV (Webster *et al.*, 1992; Lamb, Krug and Knipe, 2001; Fields, Knipe and Howley, 2007). Segment 5 encodes the major structural component of orthomyxoviruses which interacts with all of the segments to make up the complete virion (Gultyaev *et al.*, 2016).

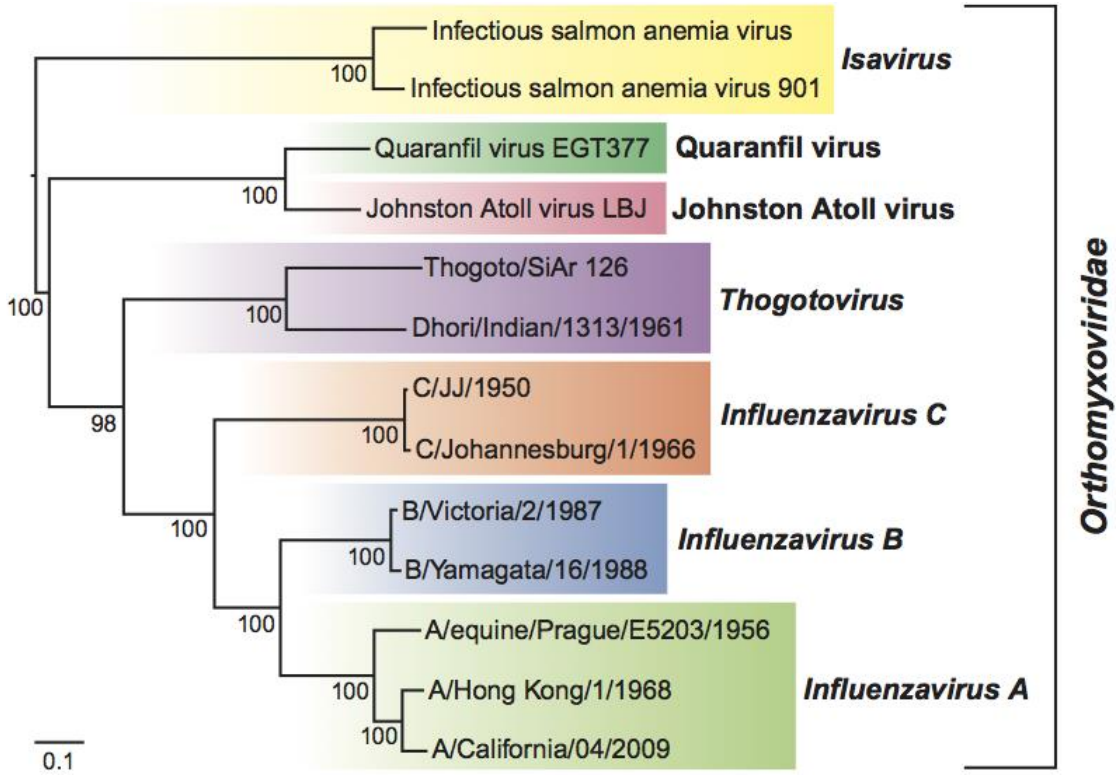


FIGURE 1.2.0 PHYLOGENETIC RELATIONSHIPS AMONG VIRUSES OF THE FAMILY ORTHOMYXOVIRIDAE GENERATED FROM NUCLEOTIDE SEQUENCES OF THE POLYMERASE BASIC 1 PROTEINS (PB1)

REPRESENTATIVES OF EACH OF THE SIX GENERA OF ORTHOMYXOVIRUSES HAVE A COMMON EVOLUTIONARY ANCESTRY.

Source: https://talk.ictvonline.org/ictv-reports/

1.2.1 INFLUENZA A VIRUSES

Influenza A virus is one of major genera within the family *Orthomyxoviridae* (Lamb, Krug and Knipe, 2001). The Influenza A virus genome is organized into eight segments 1(PB2), 2(PB1), 3(PA), encode the polymerase complex, 4(HA) and 6(NA) encode the surface glycoproteins, 5(NP) encodes the ribonucleoprotein, 7(M1/M2) encode the matrix protein and 8(NS1/NS2) encodes a nonstructural protein (Figure 1.2.1) (Szewczyk, Bienkowska-Szewczyk and Król, 2014) . Influenza A viruses have wide host ranges including humans, avians, bovines, equines and porcines (Webster *et al.*, 1992). This predisposes humans to the effects of host jumps i.e. the threat that a strain infecting non-human hosts may acquire the ability to infect humans. Such host jumps may lead to pandemics (Morse *et al.*, 2012). Influenza A viruses have the highest degree of diversity of all the known orthomyxovirus genera and are further classified based on the two major surface glycoproteins (Hoft and Belshe, 2004). Current surveillance efforts indicate that 16 HA and 9 NA subtypes of Influenza have been characterized (Poon *et*

al., 2016). New strains of Influenza virus keep emerging as is evidenced by recent infections of humans with avian and swine strains of Influenza A viruses (Smith, Vijaykrishna, *et al.*, 2009; Amonsin *et al.*, 2010; T. T.-Y. Lam *et al.*, 2013; Worobey, G.-Z. Han and Rambaut, 2014). Notable influenza subtypes that have been in circulation in humans are H1N1, H2N2, H3N2 (Scholtissek *et al.*, 1978a; Nelson *et al.*, 2008). All of these subtypes have been associated with Influenza pandemics in the last century (Lindstrom, Cox and Klimov, 2004; Rahnama and Aris-Brosou, 2013; Worobey, G. Han and Rambaut, 2014). Other Influenza subtypes that infect humans occur as a result of zoonotic transmission when non-human influenza A viruses acquire the ability to infect humans (Christman *et al.*, 2011; Lam *et al.*, 2012; T. T.-Y. Lam *et al.*, 2013; Lu, Lycett and Brown, 2014).

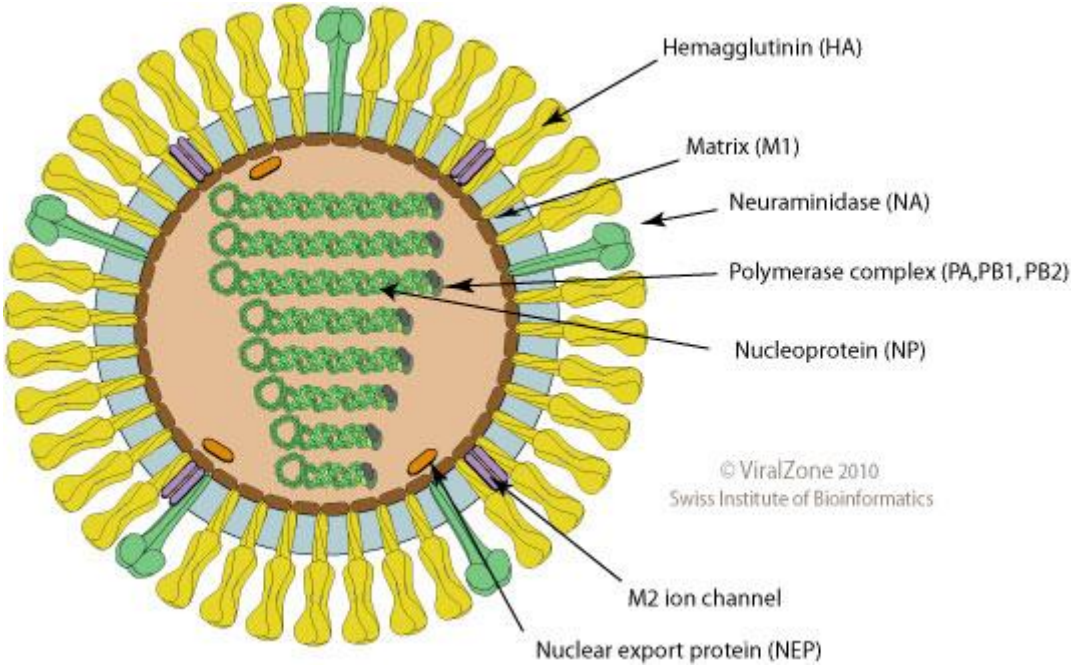


Figure 1.2.1: Illustration of Influenza A virion depicting the segment arrangement. The surface segments (HA, NA), transmembrane (M1/M2), and internal segments (NS1/NS2, PA, PB1, PB2). **Source:** **https://viralzone.expasy.org/6?outline=all_by_species**

1.2.2 INFLUENZA B VIRUSES

Influenza B viruses are one of the major genera within the family of orthomyxoviruses (Xu *et al.*, 2004). These viruses have an 8-segment negative stranded RNA genome that codes for 10-11 genes (Figure 1.2.2). It has a narrower host range compared to Influenza A viruses (Webster *et al.*, 1992; Matsuzaki *et al.*, 2004; Taubenberger and Morens, 2013). Influenza B viruses

have mostly been isolated in humans and are only second to Influenza A virus as the major etiological agent of Influenza disease in humans (Bedford *et al.*, 2015). The diversity of Influenza B virus is modest compared to Influenza A (Watson *et al.*, 2015). Before the 1980s a single strain was known but in the early 1980's it diversified into two distinct clades named after the region of their first isolation: the Victoria and Yamagata lineages (Rota *et al.*, 1990). These two lineages have continued to co-circulate and cause seasonal influenza epidemics (Hemphill *et al.*, 1993; Puzelli *et al.*, 2004; Zhao *et al.*, 2015). Global Influenza research and surveillance indicates that Influenza B is the third most prevalent cause of influenza infections after Influenza A/H1N1 and A/H3N2 (Tramuto *et al.*, 2016). The two lineages have alternatively risen and fallen from dominance between the early 1990's to the present (Oong *et al.*, 2015; Vijaykrishna, Edward C Holmes, *et al.*, 2015).

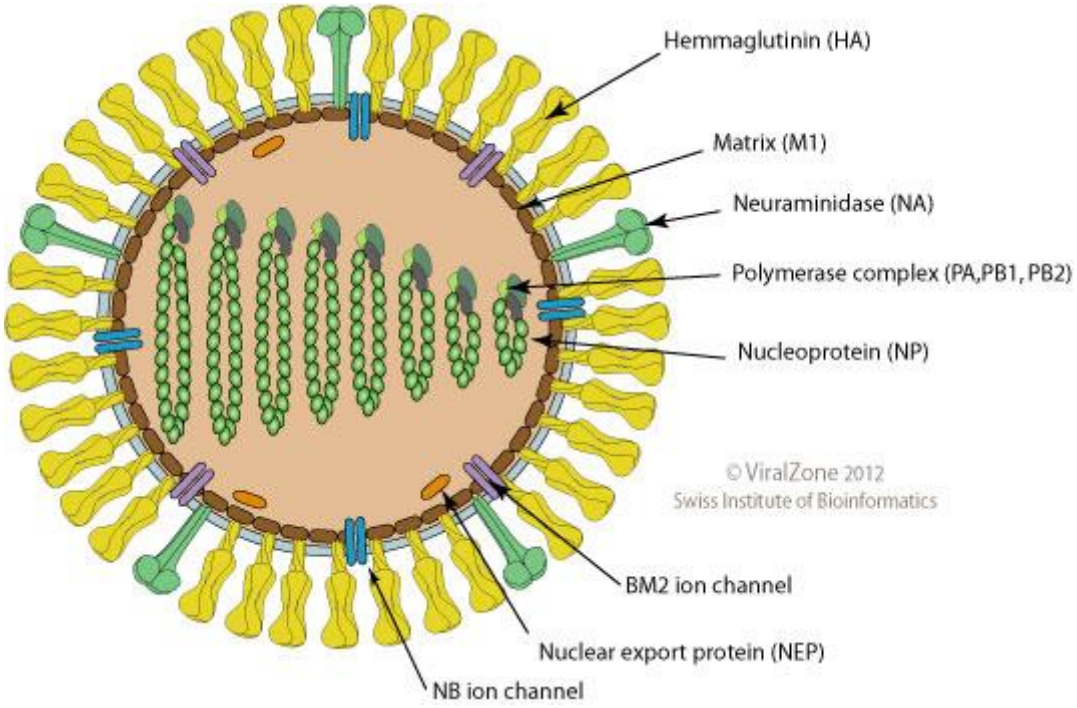


FIGURE 1.2.2: INFLUENZA B VIRUS VIRION STRUCTURE ILLUSTRATING THE SEGMENT ARRANGEMENT

The eight segments are encapsulated in capsid containing both surface and internal segments.

Source: **https://viralzone.expasy.org/80?outline=all_by_species**

1.2.3 INFLUENZA C VIRUSES

Influenza C virus is the third genus within the orthomyxovirus family. It was first isolated in 1947 and, although it had features in common with the already known Influenza A and B

viruses, it was different enough that it was classified in a genus of its own (Matsuzaki *et al.*, 2003; Odagiri *et al.*, 2015). It differs with Influenza A and B virus in that its genome comprises 7-segments. The two surface proteins, HA and NA, of Influenza A and B viruses are replaced by only one in Influenza C: hemagglutinin esterase (HEF) (Figure 1.2.3). Its genome size is estimated approximately 11.5kb. Initial epidemics were reported mainly in Japan (Compans and Oldstone, 2014; Bedford *et al.*, 2015). Influenza C is largely a human virus although very limited data shows that it can also infect other hosts such as dogs (Lamb, Krug and Knipe, 2001; Fields, Knipe and Howley, 2007). Influenza type C virus causes mild upper respiratory illnesses that remain rarer than infections associated with Influenza A and B viruses (Compans and Oldstone, 2014). One study reported six lineages through analysis of HE gene sequences. These were C/Taylor, C/Mississippi, C/Aichi, C/Yamagata, C/Kanagawa, and C/Sao Paulo. In the same study using internal genes only (PB2, PB1, P3, NP, MP and NS), two major lineages were identified: C/Mississippi/80-related lineage and C/Yamagata/81-related lineage (Matsuzaki *et al.*, 2016). Reassortment between these lineages has occurred throughout the sampling period of approximately seven decades (Peng *et al.*, 1994; Matsuzaki *et al.*, 2003, 2016; Speranskaya *et al.*, 2012).

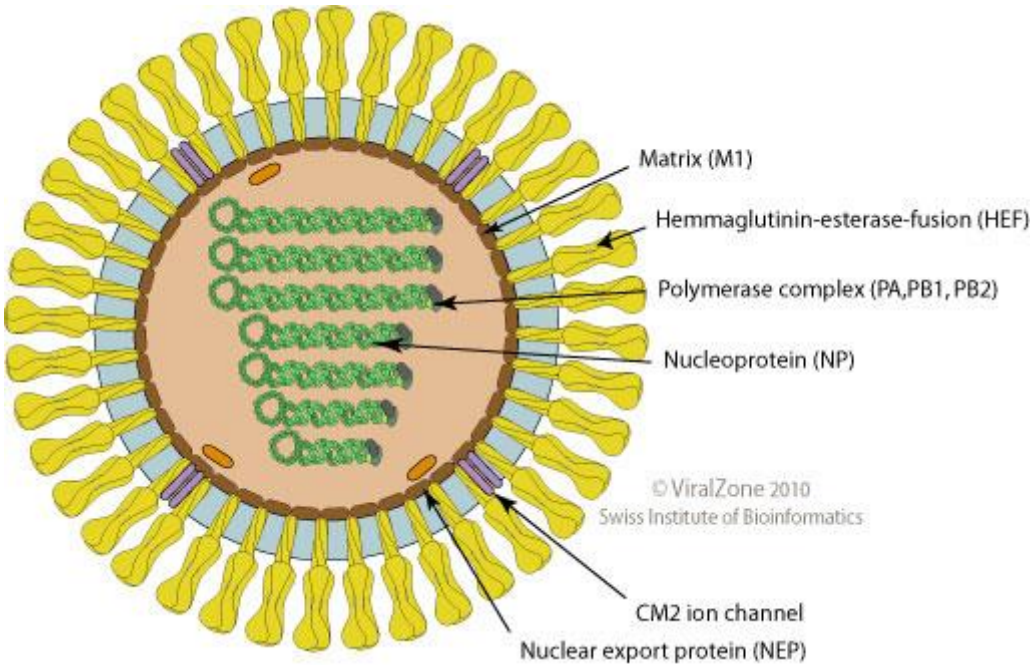


FIGURE 1.2.3: INFLUENZA C VIRUS VIRION STRUCTURE ILLUSTRATING THE SEGMENT ARRANGEMENT

HEF is the major surface protein that performs both the function of their homologous segment HA and NA in Influenza A and B virus

Source: https://viralzone.expasy.org/81?outline=all_by_species

1.2.4 Infectious Salmon Anemia Virus (Isavirus)

Infectious Salmon Anaemia virus is the fourth orthomyxovirus genus (Mjaaland *et al.*, 1997; Aamelfot, Dale and Falk, 2014; Aamelfot *et al.*, 2015). It has 8-segments and was first isolated in farmed salmon in Norway in 1984 (Figure 1.2.4)(Falk *et al.*, 1997). Since then the virus has spread to various parts of the world with outbreaks having been reported in Chile, Canada, Scotland and the United States(Godoy *et al.*, 2008) (Kibenge *et al.*, 2004; Cottet *et al.*, 2011). There are two major lineages of Isavirus in circulation; the European lineage called genotype I and the North American lineage called genotype II (Kibenge *et al.*, 2009, 2016; Mardones *et al.*, 2014). Within each genotype, there exists a spectrum of clades in circulation. Further epidemiological classification is based on a highly polymorphic region (HPR) within segment 6 that codes for th haemagglutinin esterase (HE) protein. For example strains designated as HPR0 and HPR00 are non-pathogenic to salmon and cannot be cultured whereas those designated as HPR1, HPR2 and HPR3 are virulent and cultivable in cell line cultand can be cultured within cell lines (Gagné and LeBlanc, 2017). In contrast with orthomyxoviruses that infect mammals and are cleared within a month by host immunity, Isavirus remains in its fish host for long periods (Mjaaland *et al.*, 1997). The segment orders in Isavirus also differs from those of Influenza viruses in that PB2, PB1, NP, PA, F, HA, NS1/NS2 and M1/M2 are designated as being encoded on segments 1, 2, 3, 4, 5, 6, 7 and 8 respectively (Mjaaland *et al.*, 1997).

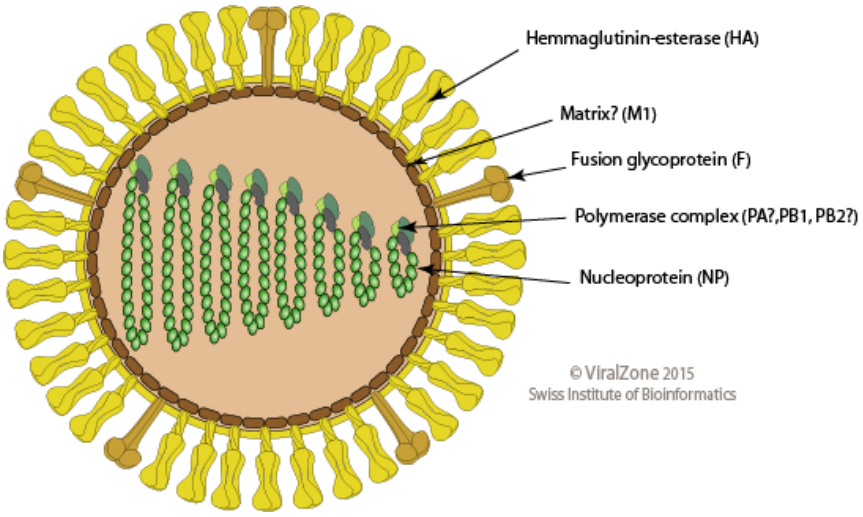


Figure 1.2.4: Infectious Salmon Anemia Virus (Isavirus), virion structure illustrating the segment arrangement. In the figure, Polymerase complex (PA, PB1 and PB2), surface glycoproteins (F, HA), transmembrane matrix (M1), nucleoprotein (NP), and the nonstructural protein (NS)-not shown

Source:https://viralzone.expasy.org/95?outline=all_by_species

1.2.5 THOGOTO VIRUS (THOV)

Thogoto virus is an arboviruses that is vectored by ticks (Michael B. Leahy *et al.*, 1997) . It is the type member of the Thogotovirus genus within orthomyxovirus family. It consists of six distinct viruses that are transmitted by both hard and soft ticks (Kuno *et al.*, 2001). These include Thogoto virus (THOV), Dhori Virus (DHOV), Batken virus (BATV) and Araguari virus (ARAV) (Peng *et al.*, 2017). They are known to cause diseases in humans especially when infections arise from Thogoto and Dhori viruses (Kosoy *et al.*, 2015; Villinger *et al.*, 2017). Thogoto viruses are distributed globally as surveillance reports indicate that they have been isolated from ticks from almost all world regions; although the prevalence may differ among the geographical continents (Lamb, Krug and Knipe, 2001; Kosoy *et al.*, 2015).

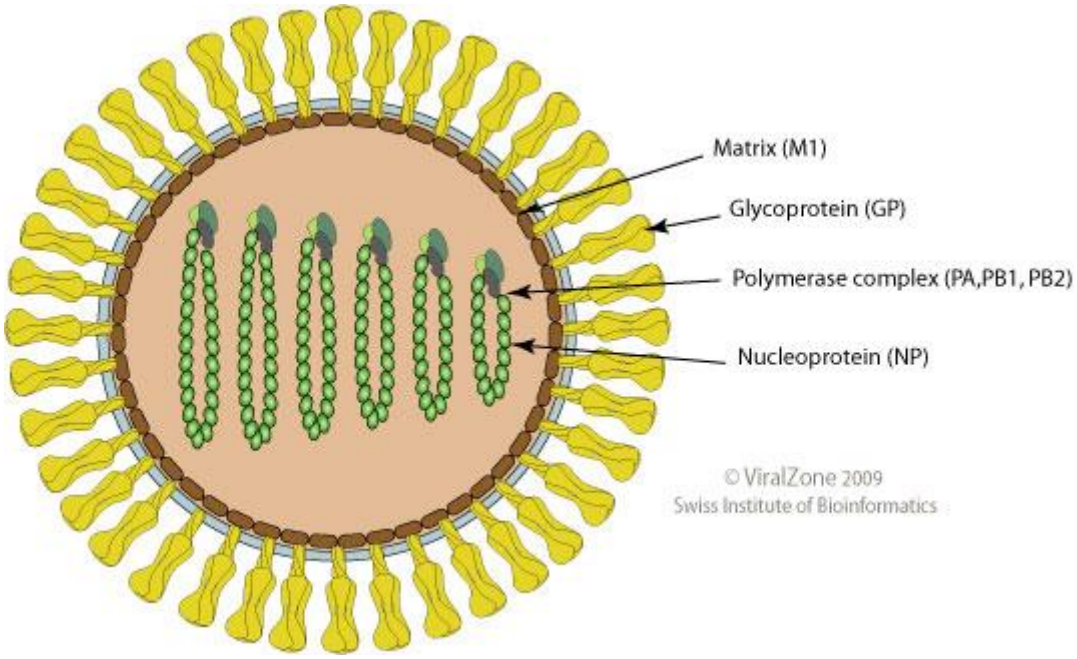


FIGURE 1.2.5: THOGOTO VIRUS VIRION ILLUSTRATING THE SIX-SEGMENT ARRANGEMENT AROUND THE NUCLEOCAPSID OF THE VIRUSTHE POLYMERASE COMPLEX CONSISTING OF PA, PB1 AND PB2, SURFACE GLYCOPROTEIN NAME GLYCOPROTEIN (GP), RIBONULEOPROTEIN (NP), AND THE TRANSMEMBRANE MATRIX PROTEIN (M1).

Source: https://viralzone.expasy.org/79?outline=all_by_species

1.2.6 QUARANJAVIRUSES

Quaranjavirus assignment as a genus within Orthomyxoviridae was proposed by Prestiet et al in 2009 after genetic and serological characterisation showed these viruses are closely related to Influenza viruses (Presti *et al.*, 2009). Viruses in this genus include Quaranfil virus, Lake Chad Virus and Jonston Atoll virus. These viruses have smaller genomes than most prthomyxaviruses (~11.5kb), spanning six segments (Presti *et al.*, 2009).

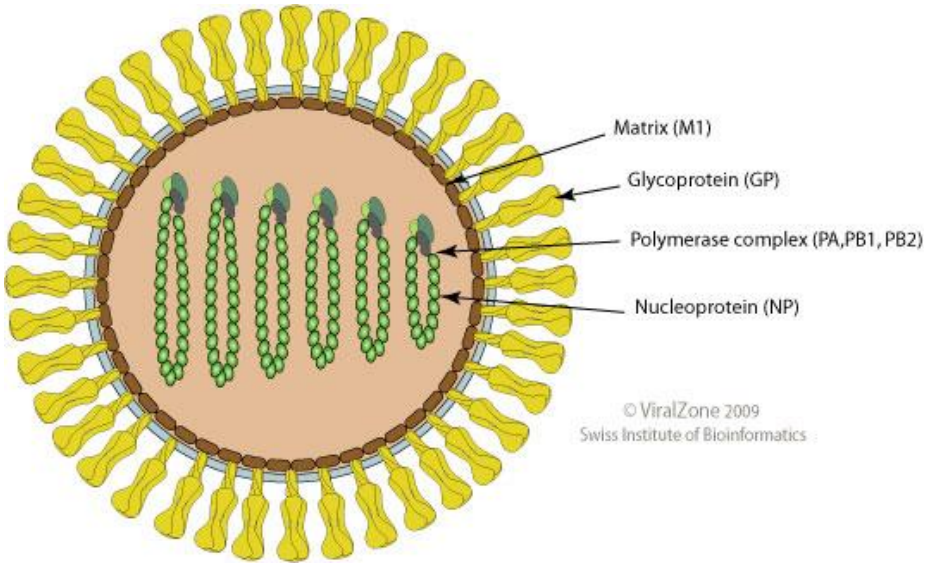


FIGURE 1.2.6: VIRION STRUCTURE AND SEGMENT MAP OF QUARANFIL VIRUS A SPECIES WITHIN THE GENUS QUARANJAVIRUSILLUSTRATED ARE THE POLYMERASE COMPLEX SEGMENTS (PA, PB1, PB2), SURFACE GLYCOPROTEIN (GP), TRANSMEMBRANE MATRIX (M1) AND NUCLEOPROTEIN (NP).

Source: https://viralzone.expasy.org/4696?outline=all_by_species

1.3 GENOMIC ORGANISATION

Influenza A and B viruses genomes span eight segments each whereas Influenza C virus contains seven segments (Webster *et al.*, 1992). Similar to Influenza A and B viruses, Isavirus has eight segments while viruses within Quaranja and Thogoto virus genera have genomes divided into six distinct segments (M B Leahy *et al.*, 1997; Kuno *et al.*, 2001; Mérour *et al.*, 2011). The genome arrangement enables the viruses in this family to perform various functions and undergo several key mechanisms that enable them to survive hostile host environments (Trifonov, 1997; Zemora *et al.*, 2016). For instance, the surface proteins are adapted to infecting cells through surface molecules (specific regions of the segment) that act as host cell receptors that facilitate virus attachment to host cells and initiate the host infection process (Shi *et al.*, 2010).

The internal segments also play crucial roles in the replication and production of viable virion (Kobasa and Kawaoka, 2005; Joseph *et al.*, 2017). The surface segments are generally more mutable than the internal segments which interact infrequently with host cells and are pressured by immune responses to change their antigenicity (Shi *et al.*, 2010). The order of segment numbering is according to the sizes of the segments - from largest to smallest.

1.3.1 MEMBRANE PROTEIN ENCODING SEGMENTS

These are segments that encode proteins found on the surface of the virion which interact with host cells and thus drive the initial entry of viruses into host cells (Puzelli *et al.*, 2004; Kobasa and Kawaoka, 2005). They include segments encoding HA, NA and M1/M2 in Influenza A and B viruses and HE and F and HA in influenza C and Isavirus (Webster *et al.*, 1992; Forrest and Webster, 2010). Thogoto virus has a GP gene that is homologous to other surface segments in other orthomyxoviruses genera (Kosoy *et al.*, 2015). Although the HA and NA are each encoded on a separate segment, M1/M2 is encoded by segment seven and undergo frame shift splicing to for the two distinct proteins: M1 and M2 (Reid *et al.*, 2002). The HA encoding segment in orthomyxoviruses plays an important role in infection. The HA protein has two active subunits, HA1 and HA2, that are proteolytically synthesized from its precursor form HA0. The process of converting to its active subunits is facilitated by host cell proteases (Neumann and Kawaoka, 2015).

1.3.2 INTERNAL SEGMENTS

These are segments that encode proteins that are not found on the membrane of the virus. They include the segment encoding the polymerase complex (PB2, PB1, PA/PE, and NS1/NS2) (Jardetzky and Lamb, 2004; Fields, Knipe and Howley, 2007). The nucleoprotein (NP) is a major structural component that interacts with other gene segments to form ribonucleic amalgamations that get packaged into infectious virions (Burkhardt *et al.*, 2014). Viral RNA polymerase is encoded by the three largest segments of orthomyxovirus genomes namely the segments encoding PB2, PB1 and PA. These encoded proteins form structural components of the virus by their interactions with ribonucleoprotein (RNP) (Martín-Benito and Ortín, 2013; Zheng and Tao, 2013). Segment eight is the smallest segment and is referred to as the nonstructural segment . The NS1 and NS2 proteins that are encoded by this segment are translated from alternatively spliced RNA transcripts (Taubenberger *et al.*, 2001). NS1 is useful for host infection whereas NS2 (which is also known as NEP) functions in the transport of newly synthesized RNPs from nucleus to the cytoplasm (Cros and Palese, 2003).

1.4. EVOLUTIONARY MECHANISMS THAT IMPACT THE GENETIC DIVERSITY OF ORTHOMYXOVIRUSES

Orthomyxoviruses like many fast-evolving RNA viruses, experience both ecological and inherent evolutionary constraints (Duke-Sylvester, Biek and Real, 2013). Several mechanisms contribute to the orthomyxovirus evolution, and the observed diversity in orthomyxovirus species is a consequence of a complex interplay between these mechanisms (Holmes and Grenfell, 2009). Specifically, the long trem

evolutionary survival of orthomyxovirus species is maintained through mechanisms such as antigenic drift, antigenic shift, point mutation and reassortment (Bhatt *et al.*, 2013). Additionally, orthomyxavirus genomes may form secondary structures that impact their evolutionary rates (A P Gultyaev *et al.*, 2014; Alexander P Gultyaev *et al.*, 2014).

1.4.1 POINT MUTATIONS

Mutations represent random errors during genome replication in living organisms (White, Enjuanes and Berkhout, 2011). As opposed to DNA viruses or higher organisms, RNA viruses possess RNA polymerase that is employed in genome replication (Taubenberger *et al.*, 2005). The mechanism of replication of fast evolving RNA viruses is inherently error prone due to lack of proof-reading activity the low fidelity of the RNA polymerases that are responsible for replication (Holmes and Grenfell, 2009). It estimated that one replication error occurs in almost every round of replication (Ueda *et al.*, 2008; Resa-Infante *et al.*, 2011; White, Enjuanes and Berkhout, 2011) mutations accumulating over large numbers of replication cycles can yield viral variants or strains with novel genetic characteristics. In Influenza viruses, the continuously accumulating mutations are classified into two major types: antigenic drifts and antigenic shifts (Bedford *et al.*, 2014; Neher *et al.*, 2015). These involve segments encoding surface proteins that interact with host derived antibodies and enable the virus to evade population-wide host immune containment (Kellam *et al.*, 2003). Antigenic shift and drift have been widely studied especially in the context of Influenza viruses (Sun *et al.*, 2013; Bedford *et al.*, 2014; Neher *et al.*, 2015; Suptawiwat *et al.*, 2017).

1.4.1.1 ANTIGENIC DRIFTS

Antigenic drifts involve the cumulative accumulation of mutations that persist over time and lead to emergence of new strains that circumvent herd immunity. Antigenic drifts are exhibited by both Influenza A and B viruses. These antigenic changes are encoded by the HA and NA segments and lead to seasonal epidemics. The greatest impact of antigenic drift is the economic burden on sustained pre-epidemic vaccination of individuals to minimize morbidity and mortality. To counter the effects of antigenic drif, vaccines against Influenza viruses have to be reformulated every year to account for the antigenic diversity of the recently circulating strains.

1.4.2 REASSORTMENT AND RECOMBINATION

Reassortment is a process in which viruses with segmented genomes randomly swap their segments leading to exchange of genetic information (Nelson *et al.*, 2008; Lam *et al.*, 2011; Viboud *et al.*, 2013). It occurs in circumstances where two or more distinct viruses with homologous segments simultaneously infect a cell. It is during replication within the host

cell,that non-random exchange of homologous segments occur between the replicating viruses. It is one of the main mechanisms that drive genetic diversity in Influenza viruses (Matsuzaki *et al.*, 2003; Nelson *et al.*, 2008; Ramey *et al.*, 2010; Greenbaum *et al.*, 2012). Increase in genetic diversity may impact on the virulence and transmission dynamics of these viruses (Neverov *et al.*, 2014; Maljkovic Berry *et al.*, 2016). The most important consequence of reassortment has been the generation of novel Influenza virus strains (Fuller *et al.*, 2013). In some cases novel viruses arising through reassortment emerge to cause regional and global pandemics (Scholtissek, 1994; Re, 2005; Taubenberger and Morens, 2006; Schrauwen and Fouchier, 2014). As with Influenza viruses, reassortment is also a significant process generating diversity in other orthomyxoviruses (Castro-Nallar *et al.*, 2011; Kosoy *et al.*, 2015; Kibenge *et al.*, 2016; Gagné and LeBlanc, 2017).

Recombination on the other hand involves non-random mixing of genetic material from within and between strains infecting a particular host (Boni *et al.*, 2010; Truong, Van Than and Kim, 2014; Chen *et al.*, 2016). In contrast to reassortment, recombination, may not involve exchange of whole segments but any region or part of a segment may also be exchanged. Also recombination may occur between homologous and non-homologous genomic regions, a departure from reassortment that requires matching segments for it to occur (Worobey and Holmes, 1999; Greenbaum *et al.*, 2012; Marshall *et al.*, 2013). Recombination in segmented viruses is thought to be rare although there are methods such as RDP4 (Martin *et al.*, 2015) that may accurately and precisely distinguish between a recombination and reassortment event within viral genomes. Recombination events detected in Influenza sequence data have generally been attributed to sequencing errors rather than from recombination *per se* (Nelson and Holmes, 2007; Boni *et al.*, 2010; Freire, Iamarino, Soumaré, Faye, Sall, Guan, *et al.*, 2015).

Interspecies transmission of Influenza and other orthomyxoviruses is frequently associated with reassortment and therefore efforts that aim at continuous surveillance and genetic characterization of circulating viruses in each ecological niche are warranted (Joseph *et al.*, 2017). Besides reassortment, host jumps may be influenced by a number of factors that including both ecological and climate changes and human-animal interactions (Peng *et al.*, 1994; Christman *et al.*, 2011; Than, Baek and Kim, 2013; Larison *et al.*, 2014; Zhu *et al.*, 2015).

A number of methods to study reassortment and recombination in segmented viruses exist (Nagarajan and Kingsford, 2011; T. T. Y. Lam *et al.*, 2013; Lu, Lycett and Brown, 2014; Martin

et al., 2015). Earlier methods involved experimental coinfections of cells with genetically distinct viruses and assessment of the extent of reassortment (Wille *et al.*, 2013; Zeldovich *et al.*, 2015; White, Steel and Lowen, 2017). With the advent of rapid sequencing technologies, abundant sequence data is now generated within reasonably short times (Metzker, 2010; Wu *et al.*, 2014). This has led to the development of computational tools that can detect genomic admixture in virus genomes with short turn-around times (Nagarajan and Kingsford, 2011; Lycett *et al.*, 2012; Westgeest *et al.*, 2014) .

The impact of reassortment is yet to be fully appreciated as most studies have focused on unraveling the genetic history of newly emergent pandemic virus after an outbreak (Lindstrom, Cox and Klimov, 2004; Taubenberger *et al.*, 2005; Galiano *et al.*, 2011; T. T.-Y. Lam *et al.*, 2013). It therefore calls for continuous assessment of circulating viruses to ascertain their changing dynamics, to understand their genetic history, to improve mitigation efforts and to control their spread in case of an outbreak (Radin, Katz, Tempia, Nzussouo, *et al.*, 2012; Hannoun, 2013).

1.4.2.1 ANTIGENIC SHIFTS

Antigenic shifts occur because of reassortment, a process during which two or more strains of Influenza virus infect a cell and a random swapping of segments might take place resulting in novel progeny viruses with new segment combinations (Xu *et al.*, 2004; Nagarajan and Kingsford, 2011; Fuller *et al.*, 2013; Vijaykrishna, Mukerji and Smith, 2015). Antigenic shifts are usually sudden and major changes in Influenza virus segments. The new virus if viable could then spread easily among hosts and can even sometimes aquire the capacity for cross species transmission. Novel viuses resulting from reassortment and a subsequent host jump will frequently emerge of new pandemic strains (Scholtissek, 1994; Re, 2005; Taubenberger and Morens, 2013; Schrauwen and Fouchier, 2014). Many past Influenza epidemics and pandemics have been associated with major antigenic shifts (Lindstrom, Cox and Klimov, 2004; Smith, Bahl, *et al.*, 2009; Vijaykrishna *et al.*, 2010; Biggerstaff *et al.*, 2014; Worobey, G. Han and Rambaut, 2014). Therefore, while antigenic drift occurs gradually over time, antigenic shifts occur only occasionally but happen almost instantly (Scholtissek, 1994; Chen and Holmes, 2008; Faria *et al.*, 2013; Urbaniak and Markowska-Daniel, 2014).

1.4.3 RNA FOLDING PATTERNS IN ORTHOMYXOVIRUSES

RNA molecules are known to naturally form relatively stable secondary and tertiary structures both *in vivo* and *in vitro* (Buratti and Baralle, 2004; Leamy *et al.,* 2016). Prediction of secondary structures within RNA virus genomes is therefore an area of active research (Gultyaev and Olsthoorn, 2010; Priore *et al.,* 2013; Alexander P Gultyaev *et al.*, 2014; Dela-Moss, Moss and Turner, 2014a) . Secondary and tertiary structures in viral genomes are thought to in many cases posess importnant biological functions..Orthomyxovirus genomes may form secondary structures that have a variety of functions (Gultyaev *et al.,* 2016; Kobayashi, Dadonaite, Doremalen, *et al.,* 2016; Lorenz *et al.,* 2016). The structural constraints on different parts of these genomes continue to be evaluated through various studies. Early studies demonstrated that in single stranded DNA viruses, DNA folding may impact on the evolution of viruses thus maintaining or preserving the viability of these viruses (Muhire *et al.,* 2014). A study using Rubella virus whole genome sequences suggested that the nucleotide substitution rates at regions of the genome that were predicted to be structured were lower that those in regions of the genome that were unstructured (Cloete *et al.,* 2014). There has been extensive investigation of secondary structure formation within Influenza viruses genomes and some of the predicted structural elements have been characterised to evaluate their impact on viral viability (Gultyaev, Fouchier and Olsthoorn, 2010; Moss *et al.*, 2012; Dela-Moss, Moss and Turner, 2014a; Kobayashi, Dadonaite, van Doremalen, *et al.*, 2016). In the current study I extend this investigation by applying computational approaches that take nucleotide sequences as input and apply analyses that combine minimum free energy and phylogenetics based approaches to predict the most likely builogically functional secondary structural elements within the genomes of Influenzavirus A, Influenzavirus B, Influenzavirus C, Isavirus and Thogotovirus.

1.5 EPIDEMIOLOGY OF ORTHOMYXOVIRUSES

1.5.1 OVERVIEW OF GLOBAL EPIDEMIOLOGY OF ORTHOMYXOVIRUSES

Orthomyxoviruses infect a wide range of hosts thus the epidemiology of different species may differ considerably (Fields, Knipe and Howley, 2007). For instance, Influenza virus epidemiology varies with type and subtype (in case of Influenza A virus) and the species of infected hosts (Rahnama and Aris-Brosou, 2013). Because of active surveillance in humans the epidemiology of influenza A in humans is better described than that in non-human hosts (Xu *et al.,* 2004; Bahl *et al.*, 2013; Vijaykrishna, Edward C Holmes, *et al.*, 2015).

The vast majority of Influenza infections are attributed to Influenza A and B viruses that cause annual epidemics, occasional pandemics and sporadic cases (Jombart *et al.*, 2009; Reperant, Kuiken and Osterhaus, 2012; Hannoun, 2013). There are also Influenza A virus subtype differences in virulence, as demonstrated in the outbreaks witnessed in the 19th, 20th and 21st

centuries (Amonsin *et al.*, 2010; Yamayoshi *et al.*, 2014; Joseph *et al.*, 2015; Zhu *et al.*, 2015). In contrast to Influenza A and B viruses, Influenza C virus cause mild respiratory disease in humans and only results in local sporadic cases (Matsuzaki *et al.*, 2003, 2016; Odagiri *et al.*, 2015). Influenza A viruses are known to have their natural reservoirs in swine and birds, and in circumstances where these "non-human" strains acquire mutations that enable them to infect humans i.e. they undergo host-switch or host-jump events pandemics or regional epidemics can occur; such as are exemplified in the 1918 Spanish flu (A/H1N1), the 2003 bird flu (A/H5N1) and the 2009 Swine-flu (A/H1N1pdm) pandemics (Tumpey *et al.*, 2005; Liang *et al.*, 2010; Christman *et al.*, 2011). Global circulation of Influenza viruses in humans is maintained by the trans-global movement of viruses within infected humans and migratory birds (Bahl *et al.*, 2013; Gardner and Sarkar, 2013; Xu, Connell McCluskey and Cressman, 2013).

Influenza B virus epidemiology is less well described than Influenza A virus epidemiology (Chen and Holmes, 2008; Bedford *et al.*, 2010; Dia *et al.*, 2013; Dudas *et al.*, 2015). Several hypotheses have been put forth to explain this discrepancy. The most striking one is that fewer Influenza B infections arise as a consequence of its genome being more stable than that of Influenza A virus i.e. its rate of evolution has been demonstrated to be much slower than that of influenza A virus (Xu *et al.*, 2004, 2014). Secondly, most influenza B nfections occur in children with naïve immune systems such that evolutionary selection pressures on the virus to evolve quickly are absent (Puzelli et al. 2004; El Moussi et al. 2013). In addition to these factors, it is suggested that Influenza B virus probably has lower receptor binding avidity for sialic acid than Influenza A virus (Vijaykrishna, Edward C. Holmes, *et al.*, 2015).

Influenza C virus epidemiology is even less understood than that of influenza B (Peng *et al.*, 1994; Odagiri *et al.*, 2015; Matsuzaki *et al.*, 2016). As with Infuenza B causes only mild respiratory disease. Although its presence has been detected worldwide, there is insufficient data available to enable researchers and policy makers to accurately quantify its impact. Epidemics associated with Influenza C virus have been reported in Japan, the USA, and the Philippines (Odagiri *et al.*, 2015; Matsuzaki *et al.*, 2016).

The epidemiology of other orthomyxoviruses such as Infectious salmon anaemia virus (Isavirus) is also not well described. Although some studies that have been conducted in regions where Salmon is reared, these have generated only limited data (Mjaaland *et al.*, 1997; Kibenge *et al.*, 2004, 2009; Godoy *et al.*, 2008; Mardones *et al.*, 2014; Gagné and LeBlanc,

2017). Isavirus is an economically important pathogen that has affected Salmon aquaculture. It remains one of the pathogens that have been prioritised for elimination in an attempt to restore the economic fortunes of Salmon farmers in those regions (Kibenge *et al.*, 2009; García, Díaz and Navarrete, 2013). Isavirus has such an important impact on farmed salmon because it has very high case fatality rates during outbreaks. Outbreaks have been reported in Canada, Chile and Norway (Godoy *et al.*, 2008; Kibenge *et al.*, 2009, 2016; García, Díaz and Navarrete, 2013; Gagné and LeBlanc, 2017). Current data on Isavirus genetic diversity suggests the existence of a European genotype and a North American genotype (Aamelfot, Dale and Falk, 2014; Kibenge *et al.*, 2016; Batts *et al.*, 2017). The European genotype has been shown to be more virulent and more genetically diverse than the North American genotype (Plarre *et al.*, 2012; Kibenge *et al.*, 2016).

Thogoto virus is transmitted by the hard and soft tick subtypes (Michael B. Leahy *et al.*, 1997). It's prevalent in Africa and southern parts of Europe (Kuno *et al.*, 2001; Kosoy *et al.*, 2015; Contreras-Gutiérrez *et al.*, 2017). Humans, domestic and wild mammals are at risk of acquiring Thogoto virus infections when they get infested with viruliferous ticks (Kuno *et al.*, 2001; Peng *et al.*, 2017; Villinger *et al.*, 2017). Dhori virus -another species within the Thogotovirus genus – circulates in parts of Asia and Western Europe (Portugal), Northern Africa (Egypt) and Russia (Michael B. Leahy *et al.*, 1997; Peng *et al.*, 2017).

1.5.2 SEASONALITY

Whereas in the temperate and arctic regions of both the northern and southern hemispheres annual winter influenza epidemics are a regular occurance, in the tropics Influenza epidemics occur throughout the year (Gessner, Shindo and Briand, 2011; Brett N Archer, Tempia, *et al.*, 2012; Katz *et al.*, 2012). In Isaviruses and Thogotoviruses there are no obvious climactically or seasonally associated patterns of outbreaks (Kibenge *et al.*, 2004, 2009; Contreras-Gutiérrez *et al.*, 2017).

The seasonality of Influenza viruses may allow transmission of circulating strains in two hemispheres as well as at the tropical regions (Forrest and Webster, 2010; Katz *et al.*, 2012; Bedford *et al.*, 2015; Neher *et al.*, 2015). These cross-regional mixing of strains impacts on the mitigation strategies (Neher *et al.*, 2015). Vaccines must be reformulated every flu season to reflect the diversity of the currently circulating strains to ensure protective and successful vaccination campaigns (Hannoun, 2013). The seasonality of influenza also aids in the control of transmission by increasing the predictability of epidemincs, making it easier for authorities

to plan for both quarantine situations and emergency measures to prevent the congregation of people in public places such as schools, markets, churches and mass transport systems (Lemey *et al.*, 2012; Antinori, Mccracken and Widdowson, 2014).

1.5.3 ORTHOMYXOVIRUS EPIDEMICS AND PANDEMICS

Several viral epidemics and pandemics in humans have been attributed to orthomyxoviruses (Gebreyes *et al.*, 2014; Metcalf *et al.*, 2015). The most striking of these have been Influenza A virus epidemics in humans, swine and birds (Lam *et al.*, 2012; Worobey, G. Han and Rambaut, 2014; Tewawong *et al.*, 2015).

Pandemics have always arisen after new Influenza virus strains have been introduced into humans from non-human hosts. Studies that have focused on the emergence of epidemics have shed useful insights into the events and circumstances that precede the generation of virus strains that lead to pandemics. Insights generated from studying previous pandemics help in formulating informed mitigation strategies for the current and future pandemics. Although there has been substantial progress in the development of methods that can improve the early detection and control of epidemics arising from infectious diseases, much still needs to be achieved, especially in developing countries in Africa Asia and Latin America (Antinori, Mccracken and Widdowson, 2014).

The 20th century Influenza pandemics were probably triggered by reassortment events between different Influenza A virus strains. In each of the pandemics either avian or swine Influenza A viruses acquired the ability to infect humans. The influenza virus that caused the 1918 Spanish flu, for example, had all segments derived from avian viruses of the A/H1N1 subtype (Taubenberger *et al.*, 1997; Morens, Taubenberger and Fauci, 2009; Worobey, G. Han and Rambaut, 2014). Similarly, the 1957 Asian flu associated with A/H2N2 virus underwent reassortment that resulted in new HA, NA and PB1 segments derived from avian hosts into humans (Scholtissek *et al.*, 1978b; Lindstrom, Cox and Klimov, 2004; Joseph *et al.*, 2015). Another Influenza A virus pandemic occurred in 1968/69 and was caused by the emergence of a novel A/H3N2 that had HA and PB1 genes of avian origin (Ortiz *et al.*, 2012). The latest epidemic emerged in early 2009 and has been suggested to have been a result of reassortment between North American classical swine A/H1N2 and Eurasian A/H1N1 lineages that formed novel swine-origin 2009 A/H1N1 pandemic virus (Viboud *et al.*, 2013; Su *et al.*, 2015a).

In contrast to Influenza A virus, other orthomyxoviruses have been implicated in local or regional outbreaks only. In addition to Influenza A/H1N1 and A/H3N2, Influenza B is a major contributor to annual winter Influenza outbreaks globally in temperate regions and causes year-round outbreaks in tropical zones (Viboud, Alonso and Simonsen, 2006; Mondini *et al.*, 2009; Antinori, Mccracken and Widdowson, 2014). Despite this seasonality, Influenza B virus has yet to cause an outbreak of pandemic magnitude. Nevertheless, the regularity of these outbreaks necessitates the trivalence of vaccines to guarantee protection from Influenza A H1N1, H3N2 and Influenza B Yam/Vic lineages (Hannoun, 2013). Recent studies have suggested that the two major Influenza B lineages have significant differences in virulence implying that both lineages (instead of just one) should potentially be included in vaccine formulations (Ampofo *et al.*, 2015).

Influenza C viruses have yet to be implicated as the etiologic agent of an influenza pandemic (Miller *et al.*, 2009; Speranskaya *et al.*, 2012). Influenza C outbreaks are generally localized to particular regions of particular countries. Although major outbreaks have been reported in Japan and the USA, globally the incidence of Influenza C remains far lower than that of Influenza A and B viruses (Matsuzaki *et al.*, 2003; Odagiri *et al.*, 2015).

Isavirus has been implicated in regional and local outbreaks in both farmed and wild salmon and related fish species in Norway, UK (Scotland), Canada, USA and Chile (Mjaaland *et al.*, 1997; Kibenge *et al.*, 2004, 2016; Godoy *et al.*, 2008; Aamelfot, Dale and Falk, 2014; Aamelfot *et al.*, 2015). The Norway outbreaks spread in a pattern that was strongly impacted by the geographical proximities of salmon farms suggesting human mediated dispersal of the virus (Gagné and LeBlanc, 2017).

Since Thogoto virus was first discovered in the 1960's in the Thogoto forest in Kenya , no major outbreaks have been reported (M B Leahy *et al.*, 1997). The virus has been detected in Africa, Asia and the USA (Kuno *et al.*, 2001; Kosoy *et al.*, 2015). A study in the USA reported new species within the Thogotovirus genus that caused a febrile infection in a male patient in Kansas state of the USA. The virus isolated from this patient was named Bourbon virus and is now one of the species within the Thogotovirus genus (Kosoy et al. 2015) . Since then Thogoto virus infections and transmissions have been only very rarely reported. Thogoto virus therefore is yet to be associated with any major epidemic or pandemic, although future active surveillance may generate data to assess the status of any future outbreaks that might be associated with thogotoviruses (Kuno *et al.*, 2001).

1.5.4 Orthomyxovirus Epidemiology in Africa with a focus on Influenza

The epidemiology of Influenza and other orthomyxovirus in Africa is still very under-studied. There is very little data available on African orthomyxovirus outbreaks than there is for outbreaks in other regions of the world (Steffen *et al.,* 2011; Katz *et al.,* 2012). Initiatives to generate more data on viruses such as Influenza A and B in Africa are still underway. For instance since the onset of the 2009 Influenza A/H1N1 pandemic, African sub-regions have either initiated or stepped-up surveillance programs aimed at describing the epidemiology of Influenza. In the next decade the actual burden of Influenza in Africa may be more accurately estimated if enough high quality data becomes available from the region (Steffen *et al.,* 2011; Radin, Katz, Tempia, Talla Nzussouo, *et al.*, 2012; Barry D. Schoub *et al.,* 2013).

Although Influenza infections occur worldwide, the burden of the disease is likely to be heaviest on people in developing countries, especially in Africa, owing to their limited healthcare programs, high frequencies of malnutrition, and the high prevalence of HIV and tuberculosis infections in many of these regions (Grassly and Fraser, 2006; Antinori, Mccracken and Widdowson, 2014). Key populations at risk of Influenza and other orthomyxoviruses include young children, the elderly, and people with weakened immune systems (such as those who are malnourished, recieving chemotherapy for cancer, taking immunosuppressant medications, or infected with viruses such as HIV which target cells of the ummune system) (Fischer *et al.*, 2014). This implies that our knowledge of influenza epidemiology in developed economies might not be directly applicable to understanding influenza epidemiology in Africa, Asia and Latin America.

In 2006 after an upsurge of avian flu H5N1 cases, the WHO, CDC and Institute Pasteur formed an initiative to strengthen preparedness of African countries to effectively identify and control outbreaks of highly virulent avian influenza viruses such as A/H5N1. This initiative, the African Network for Influenza Surveillance and Epidemiology (ANISE), aims to support the laboratory and field surveillance of Influenza cases across the African continent. These efforts have led to a more organized and planned surveillance efforts by health ministries in Angola, Côte D'Ivoire, Democratic Republic of Congo (DRC), Egypt, Ethiopia, Ghana, Kenya, Madagascar, Morocco, Nigeria, Rwanda, South Africa, Tanzania, Uganda, and Zambia (Radin, Katz, Tempia, Nzussouo, *et al.*, 2012).

The seasonality of Influenza in the temperate regions of Africa is quite well understood, this is not the case for the tropical regions of the continent (Owoade *et al.*, 2008; Fuller *et al.*, 2013;

Tempia *et al.*, 2015). Influenza in Africa is often misdiagnosed with other respiratory diseases that present with flu-like symptoms and, therefore, the burden of Influenza has likely been consistently underestimated. Further, studies that have attempted to model the dynamics of influenza in Africa have suggest that the impact of pandemic Influenza on the continent could be far greater than is currently appreciated (Bulimo *et al.*, 2012; Radin, Katz, Tempia, Talla Nzussouo, *et al.*, 2012; Theo *et al.*, 2012; Venter *et al.*, 2012; Barry D. Schoub *et al.*, 2013; Nelson *et al.*, 2014); Ortiz et al., 2012). One of these studies have suggested that a targeted study of the impact of the 2009 H1N1 pandemic in Africa should be prioritised as a means of fully reavealing the causes and consequences of influenza outbreaks in the continent (Ortiz *et al.*, 2012).

FIGURE 1.5.1 AFRICAN NETWORK OF INFLUENZA SURVEILLANCE AND EPIDEMIOLOGY (ANISE) SUB-REGIONS

Blue=North Africa; Amber=West Africa; Green=East Africa; Yellow=Central Africa; Red=Southern Africa.

1.5.4.1 THE BURDEN OF INFLUENZA IN AFRICA

Global infectious disease threats remain a serious concern. Amongst the most important of these is that posed by influenza (Fischer *et al.*, 2014). In Africa, the burden of Influenza is

poorly described (Katz *et al.*, 2012). While the southern and northern tips of the African continent both experience temperate climatic conditions and have "influenza seasons" in the winter months of the year, most of the African continent experiences tropical climatic conditions and have much more unpredictable peaks of influenza incidence (Viboud, Alonso, & Simonsen, 2006). Whereas African countries that experience temperate climatic conditions generally collect detailed data on the burden of Influenza, countries in the tropics generally lack the funding needed for the types of intensive epidemiological surveillance, diagnostic and reporting systems that are required for year-round influenza incidence tracking. It is nevertheless understood from studies in better-resourced tropical regions of the world that influenza morbidity in tropical Africa likely persists throughout the year (Viboud, Alonso and Simonsen, 2006; Antinori, Mccracken and Widdowson, 2014).

Investments in influenza monitoring in the tropical African countries are low, even accounting for the low income of these countries, because their health spending is understandably skewed in favor of more serious competing co-morbidities such as HIV, malaria and tuberculosis: all of which likely originated in Africa, and remain endemic to the African tropics and generally also have higher prevalence there than in most other regions of the world (Carter and Mendis, 2002; Wirth *et al.*, 2008; Worobey *et al.*, 2008; Vitoria *et al.*, 2009; Faria *et al.*, 2014). Influenza diagnosis requires the use sensitive PCR-based testing which is costly to implement because in addition to the high cost of individual influenza tests ($100 per test), they require governmental commitment to pay for the clinical and laboratory infrastructure (including expensive medical and laboratory personnel) that are needed to carry out these tests (Goodacre, 2013; Mason, 2016). It is particularly difficult to justify such expenses when it is considered that influenza testing will, in almost all instances, have no bearing on disease outcomes (i.e. the results of tests will only very rarely impact treatment choices) (Goodacre, 2013) .

1.5.4.2 INFLUENZA RESEARCH AND SURVEILLANCE IN AFRICA

Despite these shortcomings, a number of initiatives have been launched since the outbreak of the 2009 Influenza A/H1N1 pandemic that have focused on generating data aimed at inferring the real burden of Influenza on the African continent (Gessner, Shindo and Briand, 2011). These include Afriflu established by the World Health Organization (WHO) and the African Network of Influenza Surveillance (ANISE) established by the US Centre for Disease control and prevention (CDC), that have seen more than 15 tropical African countries establish stronger influenza surveillance systems (Steffen *et al.*, 2011; Radin, Katz, Tempia, Nzussouo, *et al.*, 2012; Barry D Schoub *et al.*, 2013; Nelson *et al.*, 2014). These efforts have come to

fruition with the support of international collaborators such as the WHO, Institute Pasteur, and the CDC that have been instrumental in providing programmatic support in terms of infrastructure and capacity building activities in the public health sectors of these countries. Importantly, besides funding substantial Influenza surveillance activities in humans, surveillance is now also being carried out in non-human hosts such as birds and pigs (Katz *et al.*, 2012).

1.6 TRANSMISSION DYNAMICS OF VIRAL INFECTIOUS DISEASES

Infectious disease outbreaks and spread remain major global health threats (Zappa *et al.*, 2009; Wu *et al.*, 2013; Bloom, Black and Rappuoli, 2017). High profile outbreaks in recent years have involved viruses such as Influenza viruses (Jombart *et al.*, 2009; Worobey, G. Han and Rambaut, 2014), SARS coronavirus (Leung *et al.*, 2003), West Nile virus (Amore *et al.*, 2010), Chikungunya virus (Devaux, 2012), Dengue virus (Mayer, Tesh and Vasilakis, 2017) and Rift Valley fever virus (Gray and Salemi, 2012; Freire, Iamarino, Soumaré, Faye, Sall and Zanotto, 2015; Venter, 2018). Many countries on the African continent are very susceptible to the devastating impacts of outbreaks caused by such viruses (Gessner, Shindo and Briand, 2011; Brunker *et al.*, 2012; Ortiz *et al.*, 2012).

The Influenza pandemic caused by the 2009 Influenza A/H1N1pdm strain was detected in Africa a few months after its emergence in Mexico (Brett N Archer, Timothy, *et al.*, 2012; Heraud *et al.*, 2012; Barry D Schoub *et al.*, 2013). Although the spatial diffusion patterns of these viruses in Africa remain unknown, various studies have documented their incidence and prevalence rates (Steffen *et al.*, 2011; Bulimo *et al.*, 2012; Nelson *et al.*, 2014; Snoeck O. J.; Sausy, A.; Okwen, M. P.; Olubayo, A. G.; Owoade, A. A.; Muller, C. P., 2015; Valley-Omar *et al.*, 2015).

Studying the spatial spread of infectious disease agents during past outbreaks is key to tackling the spread of these agents during future outbreaks (Bloomquist, Lemey and Suchard, 2010). New computational approaches that combine pathogen genomic nucleotide sequence data, demographic and geographic sampling location data – a field commonly known as phylogeography - enable spatial patterns of disease diffusion to be studied (P Lemey *et al.*, 2009; Talbi *et al.*, 2010; Lemey, 2012; Faria *et al.*, 2014; Nunes *et al.*, 2014; Gräf, Vrancken, Maletich Junqueira, *et al.*, 2015; Dudas *et al.*, 2017). They can provide novel insights into both the spatial diffusion characteristics of disease causing agents that emerge during outbreaks, and

how to best react to and contain these agents once outbreaks are detected (Nunes *et al.*, 2012; Faria *et al.*, 2013; Lemey *et al.*, 2014; Magee *et al.*, 2014; Carvalho *et al.*, 2015).

1.6.1 Molecular evolution and phylodynamics approaches to understand transmission dynamics of viral infectious diseases

Viruses with RNA genomes tend to have high evolutionary rates that enable them to adapt to new habitats rapidly (Acevedo, Brodsky and Andino, 2013; Belalov and Lukashev, 2013; van Hemert, van der Kuyl and Berkhout, 2016). The impacts of high evolutionary rates are manifested in the evasion of immune responses and development of drug resistance (Bedford *et al.*, 2010; Castro-Nallar *et al.*, 2012; Lin and Gallay, 2013; Da Costa *et al.*, 2015). Phylodynamics approaches enable researchers to gain insights into the viral infectious disease epidemiology (Takebe *et al.*, 2010; Lin *et al.*, 2011; Zehender *et al.*, 2013; Liu *et al.*, 2015; Su *et al.*, 2015a; Faria *et al.*, 2016; Dudas *et al.*, 2017). This is made possible through a combination of demographic, epidemiological and viral movement dynamics that are encoded in the topologies of phylogenetic trees reconstructed from molecular sequences (Faria *et al.*, 2013; Sessions *et al.*, 2013; Kühnert *et al.*, 2014; Vijaykrishna, Edward C. Holmes, *et al.*, 2015). Additionally, these approaches provide an avenue through which phylodynamics factors can be determined and subsequently assessed to determine their relative contributions to the virus' transmission dynamics (Lemey *et al.*, 2014; Magee *et al.*, 2014; Nunes *et al.*, 2014; Dudas *et al.*, 2017). Ultimately, they can help reveal, for any newly emerged virus, the determinants (predictors), patterns and most probable dynamics of their dissemination. Information gained from such analyses could be useful for the evidence-based formulation of mitigation, management and control strategies (Lemey *et al.*, 2012; Magee *et al.*, 2014; Gräf, Vrancken, Junqueira, *et al.*, 2015).

1.6.2 Exploring the Transmission Dynamics of Viral Infectious Diseases using viral genetic sequences

The advent of pathogen genotyping that is rapid enough that it can be done during outbreaks has enabled use of genotypic data to inform mitigation decisions and outbreak containment strategies (Assiri *et al.*, 2013; T. T.-Y. Lam *et al.*, 2013; Gire *et al.*, 2014; Faria *et al.*, 2016). Further, the availability of bioinformatics approaches that can accurately and efficiently extract epidemiologically relevant information from pathogen genomic sequence data means that pathogen evolutionary dynamics that occur during ongoing epidemics can now be monitored

in real-time (Cottam *et al.*, 2008; Grad and Lipsitch, 2014; Kao *et al.*, 2014; Sintchenko and Holmes, 2015; Campbell *et al.*, 2018).

Even in the absence of genomic sequence samples taken at the start of the epidemic, these analysis techniques can identify where and when epidemics most likely began (Studies, 2008; Takebe *et al.*, 2010; Viboud *et al.*, 2013; Xu, Connell McCluskey and Cressman, 2013). They can additionally identify likely routes of pathogen spread and, if sampling is dense enough, actual transmission chains (Grad and Lipsitch, 2014). Most importantly, however, these techniques can even provide reasonably credible estimates of key epidemiological parameters such as R_0 (de Silva, Ferguson and Fraser, 2012; Klepac *et al.*, 2014; Tuncer and Le, 2014; Metcalf *et al.*, 2015).

The continued development of computational approaches maximises the impact of pathogen genomic sequence data is already being applied to disease control and treatment. For example, the use of sequences from different circulating Influenza strains by genetic typing has informed continued vaccine formulation to that these vaccines have maximum protective effect (Hannoun, 2013; Sun *et al.*, 2013; Byrd-Leotis *et al.*, 2015; He and Zhu, 2015) .

Traditionally, case counts from clinical diagnosis have yielded disease incidence data that was used to make inferences about where and how infectious diseases spread (Tizzoni *et al.*, 2012). This kind of data in epidemiology is often referred to as line lists and includes the demography of infected individuals, their exposure, and the clinical features of the diagnosed disease (Rasmussen, Ratmann and Koelle, 2011; Norström, Karlsson and Salemi, 2012; Kao *et al.*, 2014; du Plessis and Stadler, 2015).

The tremendous power of phylodynamics derives from combining line-list data with pathogen genomic sequence data (Grenfell *et al.*, 2004; Stack *et al.*, 2010; Gray and Salemi, 2012). The genetic variation in sequences isolated during an outbreak allows inferences to be made about the direction of transmission and could be corroborated with routes of transmission to describe fully the transmission dynamics which could not be possible using traditional line-lists alone (Frost *et al.*, 2015; Metcalf *et al.*, 2015) .

Genomic sequence data coupled with line-data provides an unprecedented opportunity for the investigation of medium to long-term evolutionary dynamics of rapidly evolving viruses (Sim and Hibberd, 2016). This is achieved through the application of coalescent theory which posits

that sets of sampled genomic sequences have, independent of real-time case-counts, encoded within them information on population growth or contraction and the dynamics of past transmission events (Kühnert, Wu and Drummond, 2011; Norström, Karlsson and Salemi, 2012; Gebreyes *et al.*, 2014). Within a probabilistic framework these coalescent-theory based approaches also have the capacity to, in addition to line-data, account for any amount of additional metadata such as the geographical features of the landscapes across viruses move, spatially varying features of host populations and environmental conditions (Lemey *et al.*, 2014; Dudas *et al.*, 2017).

1.7 VIRAL PHYLODYNAMICS

Phylodynamic inference based on pathogen genetic sequence data continues to be powerful supplement to epidemic surveillance (Rife *et al.*, 2017; Volz, Romero-Severson and Leitner, 2017a). It has currently become a mature field that aims to enhance our understanding of infectious disease transmission and evolution (Grenfell *et al.*, 2004). Phylodynamics relies on phylogenetic inference as the core analytical tool to recover evolutionary and epidemic processes (Baele *et al.*, 2017). This is achieved by analysis of the mutations that accumulate in the genomes of rapidly evolving pathogens during the expansion of an epidemic. These mutations may confer phenotypic differences that allow viruses to infect different cell types, to evade host immune responses or to transmit by different routes, hosts or vectors (Vijaykrishna, Mukerji and Smith, 2015), but the mutations may also represent the molecular footprint of epidemiological processes that can otherwise not directly be observed. The primary goal of phylodynamics remains extracting such information from genetic data and requires the integration of additional data and models in a phylogenetic framework (Baele *et al.*, 2017). It therefore follows that phylodynamics is not only delineating the interplay between evolution and epidemiology from a conceptual stance, but also integration is made concrete through advances in statistical modeling and computational inference (Volz, Romero-Severson and Leitner, 2017a; Baele *et al.*, 2018). Time of sampling therefore represents important information to incorporate in phylodynamic analyses because it allows calibration of phylogenies, and hence epidemic histories of rapidly evolving pathogens in calendar time units (Pybus and Rambaut, 2009; Ho *et al.*, 2011; Biek *et al.*, 2015) Molecular clock models that draw direct relationships between sequence divergence and evolutionary time have been extended specifically for this purpose, and models accommodating the time of sampling form the core theme of time-

scaled phylodynamics (Drummond *et al.*, 2005; Shapiro, Rambaut and Drummond, 2006; Metzker, 2010).

Phylodynamics approaches encompass strategies and techniques that integrate evolutionary and ecological processes of pathogens (Grenfell *et al.*, 2004; Pybus and Rambaut, 2009; Faria *et al.*, 2011). These processes have been shown to be inextricably linked for fast evolving pathogens such as RNA viruses. Consequently, these approaches were initially used to decipher transmission dynamics from genetic sequence data. However over time, there has been great development in the computational phylodynamics techniques and many more sophisticated tests can now be done on multidimensional datasets (Beard *et al.*, 2014; Baele *et al.*, 2017; Magee, Suchard and Scotch, 2017; Rife *et al.*, 2017).

Phylodynamics methods have grown in leaps and bounds due to the availability of large amounts of genetic sequence data that has increased in parallel with high throughput sequencing techniques and the ongoing development of advanced statistical techniques (Duchêne, Holmes and Ho, 2014).

Phylodynamics is easily applied to RNA viruses as they have a major characteristic of evolving rapidly in response to selection pressure exerted by host immune response and also to enable their adaptation to new habitats (Volz, Romero-Severson and Leitner, 2017a). Furthermore, viral disease control is impeded by the ability of RNA viruses to infect different hosts that may emerge as new susceptiple populations, a situation that could easily lead to major outbreaks of epidemic and pandemic scales (Engel *et al.*, 2013; Venter, 2018). Pathogen phylodynamics was traditionally used to demonstrate how biological and epidemiological processes impacted on the pathogen's phylogeny (Stack *et al.*, 2010; Neher, Russell and Shraiman, 2014; Trovão *et al.*, 2015). It is now well understood that for measurably evolving pathogens, evolutionary dynamics can be heavenly influenced by epidemiological and immunological processes (de Silva, Ferguson and Fraser, 2012; Acevedo, Brodsky and Andino, 2013; Lemey *et al.*, 2014; Lee *et al.*, 2015).

RNA viruses form some of the most important human pathogens such as HIV, HCV, Influenza, WNV and Zika viruses (Worobey *et al.*, 2008; Lewis *et al.*, 2015; Olmstead *et al.*, 2015; Venter, 2018) . Key insights that have been gleaned from phylodynamic investigations have included the timing of the emergence of epidemics , species or reservoir population in which the pandemic originated, the timing and order of transmission events of an outbreak and spatial routes of dispersal and recently key drivers of the observed dissemination pattern can now be explored precisely (Smith, Vijaykrishna, *et al.*, 2009; Worobey, G.-Z. Han and Rambaut, 2014; Olmstead *et al.*, 2015; Rife *et al.*, 2017).

1.7.1 METHODS

Phylodynamics methods and techniques have kept evolving with advances in computational and bioinformatics methods in recent times. Although initially coined by Grenfell in 2004, that the concept of phylodynamics dates as far back as 1876 when Haeckel reconstructed a phylogenetic tree using phenotypic traits in a study that attempted to describe the center of creation and the distribution of twelve races of man (Stauffer, 2004; Rieppel, 2011). The combination of phylogenetic relationship and spatial information developed into what is currently a defined branch of phylodynamics commonly known as phylogeography (P Lemey *et al.*, 2009; Bielejec *et al.*, 2011).

Most phylodynamics methods have been implemented in Bayesian inference frameworks (Drummond and Rambaut, 2007; Drummond *et al.*, 2012; Drummond and Bouckaert, 2014; Suchard *et al.*, 2018). The advantage of the Bayesian framework is that it incorporates well evaluated uncertainty when full probabilistic models are used and allows integration of time-scaled evolutionary histories of genetic sequences under alternative molecular clock models with other trait, evolutionary and demographic processes (Yang and Rannala, 1997; Baele *et al.*, 2018).

Population genetic modelling can enhance Bayesian phylogenetic inference by providing a realistic prior on the distribution of branch lengths and times of common ancestry (Nascimento, Reis and Yang, 2017) .The parameters of a population genetic model may also have intrinsic importance, and simultaneous estimation of a phylogeny and model parameters has enabled phylodynamic inference of population growth rates, reproduction numbers, and effective population size through time (Drummond *et al.*, 2005; Heller, Chikhi and Siegismund, 2013; Villinger *et al.*, 2017).

The most popular tool at the disposal of researchers applying phylodynamic methods is the computer program BEAST (Bayesian Evolutionary Analysis by Sampling Trees): a tool that has been under active development since 2001 (Drummond and Rambaut, 2007; Drummond *et al.*, 2012; Suchard *et al.*, 2018) . In addition to BEAST, there are a number of other Bayesian and maximum likelihood programs that used to perform phylogenetic tree reconstruction from genetic sequence data e.g MrBayes, PAML, PhyML, GARLI, ExaBayes, LSD, TreeTime, PhyloType, BayesTraits and several R packages(Baele *et al.*, 2018) (Figure 1.7.1)

Many of these computer programs require high performance computing helper libraries to function at peak performance on multi-core processor platforms either as Central Processing Unit (CPU) or Graphical Processing Unit (GPU) architectures. This is particularly useful in software that implements massively parallel computations eg RAxML, PhyML and ExaBayes (Aberer, Kobert and Stamatakis, 2014).

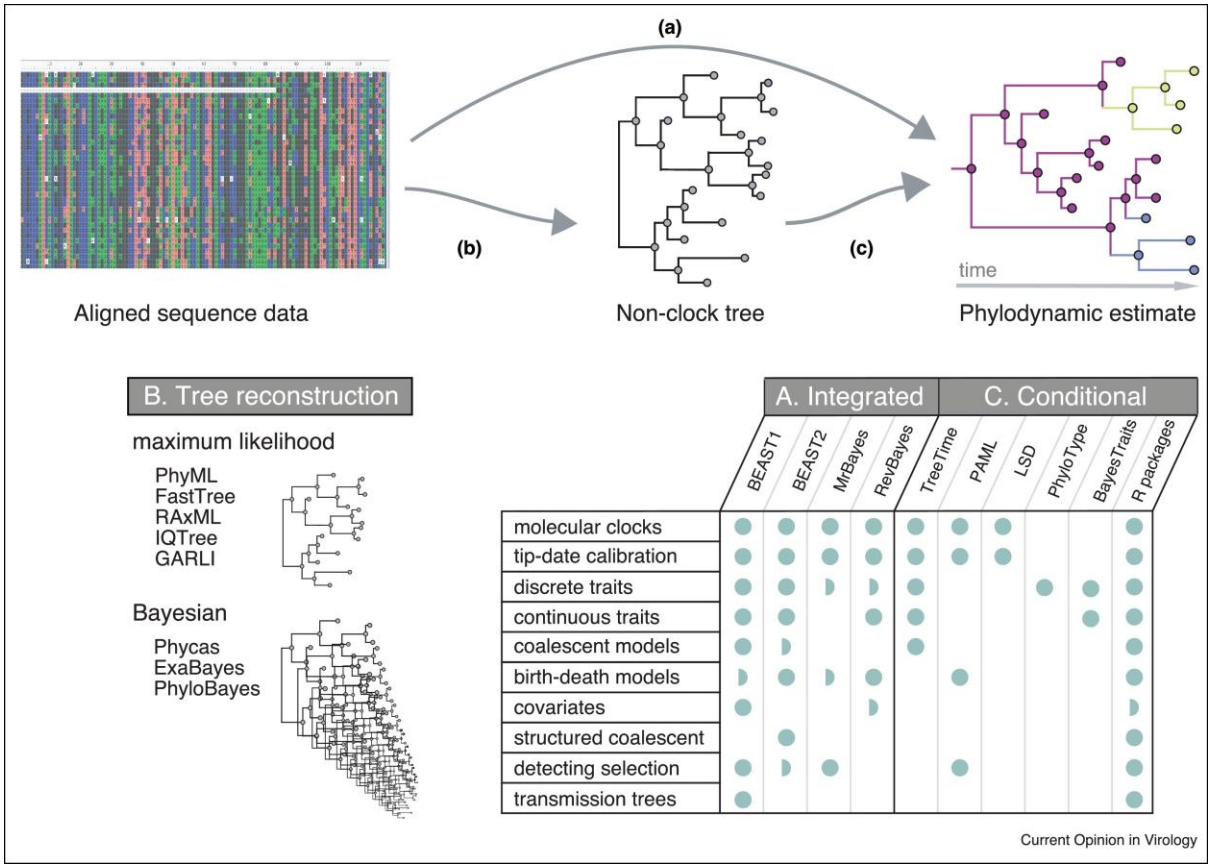
Two popular helper libraries implemented in phylodynamics analysis software packages are: Broad-platform Evolutionary Analysis General Likelihood Evaluator (BEAGLE) and Phylogenetic Likelihood Library (PLL) (Flouri *et al.*, 2015; Ayres *et al.*, 2019). BEAGLE speeds up phylogenetic calculations for existing software packages enabling more effective use of available multi-core hardware, including GPUs whereas PLL is a highly optimized application programming interface for performing likelihood-based phylogenetic inference, targeting multi-core processors such as the Intel Xeon Phi using MPI. While BEAGLE can be used with BEAST 1, BEAST2, MrBayes, PhyML and GARLI, PLL has been integrated with DPPDiv and IQ-TREE . DPPDiv estimates divergence times on a fixed tree topology under a Bayesian framework.

Although these computer programs have a common strategy of using phylogenetic trees as input to generate similar phylodynamics estimates, they implement a range of different procedures and models. A number of them can rescale branch lengths from substitutions per site into calendar time units from time-stamped sequences e.g . TreeTime, PAML and LSD. On the other hand, BayesTraits and PhyloType are more useful in analyses that seek to analyse the phylogeography of an epidemic (Pagel, Meade and Barker, 2004; Chevenet *et al.*, 2013).

Apart from the above programs which perform specific inference tasks, several packages have been developed in R that can perform similar analyses for phylodynamic inference. These include ape, phytools, treedater, rcolgem, skygrowth and phyland (Baele *et al.*, 2017; Rife *et al.*, 2017; Paradis and Schliep, 2019). More packages developed in R that perform phylodynamics inference are hosted online at https://cran.r-project.org/web/views/Phylogenetics.html and http://www. repidemicsconsortium.org/.
In tandem with the above tools, more tools have been developed to aid in visualisation of the inferences made from phylodynamic analyses such as Spatial Phylogenetic Reconstruction of EvolutionAry Dynamics (SPREAD), Microreact and Nextstrain (Bielejec *et al.*, 2011; Argimón *et al.*, 2016; Hadfield *et al.*, 2018). Microreact comes as a web application for visualizing data sets consisting of any combination of trees, geographical, temporal and associated (meta)data.

Nextstrain aims to visualise epidemics as they unfold in real time by regularly updating public databases for the latest sequences and using fast maximum-likelihood reconstructions. It consists of a database of viral genomes, a bioinformatics pipeline for phylodynamics analysis, and an interactive visualisation platform based on Python and JavaScript.

The computational and methodological advances that made phylodynamic inference possible for increasingly large data sets have also created a need to visualize large phylogenetic trees as well as any annotated information (e.g. virus and patient data). iTol and PhyloGeoTool are particular examples that address this need by supporting interactive navigation of large phylogenies and exploration associated clinical and epidemiological data (Libin *et al.*, 2017; Letunic and Bork, 2019; Theys *et al.*, 2019)



Source: Current Opinion in Virology 2018, 31:24–32

Figure 1.7.1 Schematic representation of a phylodynamic inference work flow highlighting most commonly used model frame works and software packages.

1.7.2 APPLICATIONS OF PHYLODYNAMICS

1.7.2.1 PATHOGEN ORIGINS

The most important question during an outbreak relates to researchers knowing the origin of the pathogen causing the outbreak. Origin could be inferred by considering both the timeline of emergence and geographic location (Duke-Sylvester, Biek and Real, 2013). Knowing these two factors may guide intervention or mitigation of further spread of the pathogen. The reconstruction of the origins of pathogen over long or short time scales requires calibrations of the timescale in calendar time units such as days, months or years (Gog *et al.*, 2014; Abecasis, Pingarilho and Vandamme, 2018). Sequence data whose dates of sampling is known-commonly referred to as time-stamped data form the core ingeredients of virus or any other pathogen phylodynamics. It is this kind of data that feed into molecular clock models (ie

models that describe the relationship between genetic distances and time among nucleotide sequences) (Hughes *et al.*, 2009; Tee *et al.*, 2009; Rahnama and Aris-Brosou, 2013).

Although earlier applications of phylodynamics were in rapidly evolving viral pathogens recent developments show that these methods could be applied to other somewhat slow evolving pathogens such as bacterial if long term scales are the main focus (Prosperi *et al.*, 2013; Azarian *et al.*, 2016; Rife *et al.*, 2017).

Infectious disease transmission is an inherently spatial process, hence the inference of viral geographic dispersal through phylogeography studies in Bayesian statistical framework are used to infer epidemic origins and obtain a description of the spatial spread based on either discrete or continuous location data (P Lemey *et al.*, 2009; Lemey *et al.*, 2010; Faria *et al.*, 2011). Discrete trait analysis continues to be frequently used to reconstruct migration patterns of viruses or other pathogens modelled as instantaneous movements (also referred as to 'jumps') that may occur at different frequencies between a set of observed locations that were sampled (P Lemey *et al.*, 2009). Discrete trait analysis is implemented in the same approach as continuous time markov chain models that used in modelling genetic sequence nucleotide substitutions. In the same frame work, factors that impact on the observed spread can now be assessed through Continuous Time Markov Chain (CTMC) through generalized linear modelling to evaluate their contribution to the observed spread pattern of the pathogen in question (Lemey *et al.*, 2010; Gill *et al.*, 2017). On the otherhand, continuos diffusion model approach allows reconstruction of spatial history considering both observed and unobserved sampling locations (Lemey, 2012). Modifications of these models enable flexible and realistic diffusion patterns to be described for example instead of assuming similar diffusion rate in Brownian diffusion way, rates of diffusion are allowed to vary between locations and in different directions(Lemey *et al.*, 2010; Gill *et al.*, 2017)

1.7.2.2 COALESCENT THEORY AND PHYLODYNAMICS -EFFECTIVE POPULATION SIZE

In phylodynamics, the coalescent is an important theory that helps illuminate the population dynamics of a certain lineage of pathogens through time (Drummond *et al.*, 2005; HO and SHAPIRO, 2011). The coalescent enables the population changes of present day circulating genotypes to be traced backwards in time to undersatd their growth. The coalescent model assumes that the present day population of pathogens grew in vertical manner in the absence of recombination (Pybus and Rambaut, 2009; Volz, Koelle and Bedford, 2013; Volz, Romero-

Severson and Leitner, 2017a). In Bayesian phylogenetics, the dynamic process of pathogen transmission through time translates to the shape of the time-measured phylogenies (Redelings and Suchard, 2005; Drummond and Rambaut, 2007; Drummond *et al.*, 2012; Drummond and Bouckaert, 2014; Suchard *et al.*, 2018). This forms the core assumptions of the coalescent or birth-death models used to gain quantitative phylodynamic insights from genetic sequences of fast evolving viruses (Nascimento, Reis and Yang, 2017). Continued development of the coalescent models has given rise to various flavors of the models to suit the complex dynamics that are typical of fast evolving RNA viruses among other pathogens (Rasmussen, Ratmann and Koelle, 2011; Volz, Koelle and Bedford, 2013; Ratmann *et al.*, 2017).

These approaches have made interpretation of analyses that sought to decipher viral spread in ancestral settings and evaluation of factors that may have influenced the observed transmission pattern easier than before. In the latest developments, investigators have been able to formally test and evaluate predictors of viral population dynamics through application of generalised linear models (GLMs) in the inference procedure (Faria *et al.*, 2013; Lemey *et al.*, 2014; Nunes *et al.*, 2014; Tuncer and Le, 2014; Gräf, Vrancken, Junqueira, *et al.*, 2015; Dudas *et al.*, 2017). The quest to address epidemiological questions have also received a boost through coupling of coalescent or birth-death models to complex models of infectious disease dynamics thereby building on previous works that provides seamless link between phylogenetic (tree-generative) models and compartmental models such as Susceptiple -Infected (SI), Susceptible-Infeceted-Susceptible (SIS), or Susceptible-Infected-Recovered (SIR) (Grassly and Fraser, 2006; Koelle *et al.*, 2011; Brett N Archer, Tempia, *et al.*, 2012; Grear *et al.*, 2018). Incidence and prevalence dynamics as well as tailored transmission processes can now be calculated from genetic data through implementation of more advanced statistical models such as particle filtering also known as sequential Monte Carlo (SMC) (Giardina *et al.*, 2017).

1.7.2 .3 VIRAL ADAPTATION/SELECTION

Hypotheses of selection and population structure can be assessed directly through application of methods from computational phylogenetics and population genetics following conventional evolutionary approaches . For example, the comparison of the rate of synonymous substitution and nonsynonymous substtituions (dN/dS) may reveal the level of selection in a pathogen population whereas the F statistics from genetic sequences could be used to reveal the extent of population structuring of a given viral population. These analyses when applied in the framework of coalescent theory, can guide epidemiological inferences and thus embrace the underpinnings of phylodynamics.

1.7.3 CHALLENGES AND LIMITATIONS OF PHYLODYNAMICS

Although phylodynamics as an approach to pathogen research has advanced rapidly in terms of application and methods development, it still has a number of open challenges.

1.7.3.1 ACCOUNTING FOR SEQUENCE SAMPLING PATTERNS

The sampling pattern of sequences in public databases is more often biased in the sense that sampling may be biased towards trying to capture a diverse taxonomic sample or may be biased by sampling a particular geographic area, and if such samples are used in estimating the effective population size then some adjustments would inadvertently have to be made to account for the sampling bias (Maljkovic Berry *et al.*, 2016). Phylogeography of a pathogen is heavily influenced by the sampling pattern and therefore sampling effects need to be accounted for proper inferences to be made about the spatial patterns of the pathogen (P Lemey *et al.*, 2009; Lemey *et al.*, 2010; Faria *et al.*, 2011). The problem of sampling bias in phylogeography is very important in the sense that sample locations, are treated as mutations and so if a location is oversampled then some models will increasingly invoke that location as the sink in spatial diffusion of pathogens (Karcher *et al.*, 2016).

Inferences based on coalescent models tend to have higher degree of accuracy when temporal sampling is conducted at a specific point in the epidemic cycle as demonstrated in a study performed by Stack et al., 2010 (Stack *et al.*, 2010). On the other hand, birth-death models tend to make assumptions of a constant probability of sampling throughout the evolutionary history, which may result in biased estimates of quantities such as the effective population size when the sampling process is mis-specified for phylodynamic inference (Lepage *et al.*, 2007; Kühnert *et al.*, 2014).

In order to make the best use of currently available data, methods of sampling could be selected and classified according to surveillance information; also, derivation and use of realistic models of the sampling process could in theory improve statistical power while lowering bias (Frost *et al.*, 2015; Baele *et al.*, 2017; Rife *et al.*, 2017). Potentially confounding effects of both spatial and temporal non-random sampling should be investigated and more realistic ways to explain and avoid them done.

1.7.3.2 NO REALISTIC EVOLUTIONARY MODELS

In estimating the rate of evolution various models are used that make several assumptions. In studies involving rapidly evolving viruses, the sampling times are often used to estimate the divergence times on a phylogenetic tree as they enable the calibration of molecular clocks in calendar units such as days, months or years (Drummond *et al.*, 2006; Bromham *et al.*, 2018). The molecular clocks come in two flavours; the strict clocks and relaxed clocks where the former assumes constant rate of evolution among entities whereas the later assumes that rate of evolution among taxa may vary with time. Although the relaxed clocks allow some degree of variation of the evolutionary rates, they still have a limitation as they fail to reveal the true variation in the rates of evolution (Suchard and Drummond, 2010; Fourment and Darling, 2018).

This phenomenon has been studied by researchers who reported that current molecular clock models failed to capture fine-scale temporal correlations in the evolutionary rate (Worobey *et al.*, 2016). This was shown in a study of HIV major subtypes which are thought to be closely related but have different substitution rates that each lineage had markedly different divergence times when analysed separately in contrast to when all the subtypes were combined and analysed together. Using Influenza virus sequences Worobey in 2014 found that the above problem could be reduced or solved by incorporating host species information in calibration of the molecular clock timescale (Worobey, G.-Z. Han and Rambaut, 2014).

In addition, most studies do not consider how the epidemiological dynamics may feed back to the pattern of evolution (Bedford *et al.*, 2010; Frost *et al.*, 2015). Some viruses have characteristic evolutionary mechanisms that deviate from neutral coalescent expectation resulting in asymmetric tree topologies e.g. Influenza virus evolution forms a trunk or ladder like tree where only lineages that escape herd immunity manage to spread in the population (Bedford *et al.*, 2010; Baele *et al.*, 2012; Meyer *et al.*, 2015). Models designed to analyse such evolutionary dynamics are still lacking although efforts to address this has resulted in the application of Bolthausen–Sznitman coalescent that allow for selection pressures to be accounted for (Viboud *et al.*, 2013; Brunner *et al.*, 2019). In spite of these, assumptions of these models remain unrealistic as they fail to provide a clear link between viral mutations and evolutionary outcomes and assume that the resulting phylogeny is independent of the substitution process.

1.7.3.3 INFLUENCE OF STOCHASTIC EFFECTS

In infections that result in recurrent epidemics, such as influenza A virus and norovirus, demographic stochasticity often play a key role in seasonal downward dynamics in incidence. Similarly, current endemic infections in populations such as HIV and Hepatitis C virus were at one time detected in low frequencies and may sporadically erupt in new populations (Rambaut *et al.*, 2016). These implies that stochastic effects may exert their pressure at the time when the most recent common ancestor of many pathogens existed. Stochastic effects due to demography may also be important when the number of infected individuals is relatively small and/or infection and recovery rates are high (Nicolaides *et al.*, 2012; Duke-Sylvester, Biek and Real, 2013). Many population dynamics models in viral studies assume constant rates and maybe appropriate in outbreak situations where the virus usually spreads exponentially but inappropriate non-outbreak situation when virus is not spreading exponentially (de Silva, Ferguson and Fraser, 2012). This scenario has been applied in several studies that employed linearized stochastic birth-death processes whereby birth imply transmission, whereas death is synonymous to recovery or death of the infected individuals (Rasmussen, Ratmann and Koelle, 2011).

To overcome the challenge of assumption of constant rates of population changes, modifications of the conventional models has led to rise of new models such as the birth-death skyline being developed (Lepage *et al.*, 2007; Kühnert *et al.*, 2014; Paraskevis *et al.*, 2015). This new model attempts to account for varying infection rates and involve fitting a piecewise constant birth-death process. Mechanistic insights may also be gained by fitting a stochastic nonlinear model of disease transmission rather than from the earlier described nonparametric approaches. This approach was implemented in a study by Rasmussen et al 2011 who incorporated stochasticity through application of the coalescent to a collection of stochastic simulations by fitting a stochastic differential equation model jointly to epidemiological data and to coalescent events deciphered from a phylogeny (Rasmussen, Ratmann and Koelle, 2011). Similarly, this can be achieved by allowing the use of coalescent likelihood to perform a stochastic change in timescale. For instance, this was demonstrated in a study be Kuhnert et al 2014 who parameterised a stochastic birth-death process with piecewise constant rates by fitting a stochastic epidemiological model by simulation of epidemiological trajectories (Kühnert *et al.*, 2014). Further, there are ongoing developments where incomplete data is fitted with stochastic epidemic models. Using numerical approximation to the solution of the

underlying master equation, Leventhal et al 2016 fitted a stochastic model by integrating out the (unknown) number of transmission events in the population that occurred between coalescent intervals in the sample (Leventhal *et al.*, 2016). There are also gaps in current literature of phylodynamics regarding environmental stochasticity, a factor that plays a role in understanding the dynamics of several vector-borne diseases (Worby *et al.*, 2016). Continued developments as witnessed recently where researchers performed analyses by fitting models that can accommodate both stochasticity (via stochastic differential equations, a common framework for including environmental stochasticity) and structure were encouraging and await wider availability of sequence data on vector-borne pathogens (Frost *et al.*, 2015; Rife *et al.*, 2017).

1.7.3.4 EFFECT OF RECOMBINATION AND REASSORTMENT

Reassortment and recombination may occur when host cells are simultaneously infected with two or more strains/lineages of viruses implying that pathogens potentially with different times and places of origin may be present in the infected host at the same time (Worobey and Holmes, 1999; Philippe Lemey *et al.*, 2009).

Most phylodynamics studies assume there is limited or no recombination/reassortment or apply 'multiple loci' model which assumes free recombination among sequence partitions, but no recombination within them.

The impact of selection on the shape of trees is illustrated in phylogenies reconstructed from Influenza A virus and HIV-1 glycoproteins (Kosakovsky Pond *et al.*, 2006). The resulting phylogeny which normally has ladder-like topology is the hallmark of strong directional selection as the viruses evolve to evade host immune pressure thereby resulting in an imbalanced tree. In contrast to this, a virus that is not subject to strong directional selection results in more balanced phylogenetic tree reconstructed from sequences isolated from different individuals in a population (Simon-Loriere and Holmes, 2011).

Segmented viruses are unique in the sense that they enable researchers to overcome the problem of accounting for the effects of recombination during their phylogenetic reconstructions as only the frequency of recombination events is what matters rather than the locations of the breakpoints (McDonald *et al.*, 2016). Analysing reassortment for datasets derived from different segments in segmented genome viruses is challenging and is sometimes

similar to challenges experienced when modelling host-parasite coevolution to understand how evolutionary dynamics in the host affects adaptive evolution of parasites.

On the other hand, phylogenies of the HIV envelope protein from long established infections tend to be similar to influenza's ladder-like tree. This suggests that the processes affecting viral genetic variation are non-uniform across scales. Indeed, different patterns of viral genetic variation within and between hosts his still an active topic in phylodynamic research (Santiago and Rivera-Amill, 2015).

Viruses undergo recombination and reassortment (segmented viruses) and if recombination is not considered during phylogenetic analysis, it is likely to lead to errors in phylogenetic inference. The effects of recombination are imprinted in the phylogeny and result in star-like phylogenetic trees if reconstructed from recombinant sequences. This may subsequently be erroneously treated or associated with exponential growth of the viral population. Accounting for recombination helps overcome such errors and may provide deeper insights into the transmission dynamics of the viruses (Schierup and Hein, 2000).

Approaches that explore joint inference of phylogeny and recombination are actively being developed and would be useful in addressing the problems of recombination in phylogenetic reconstruction. Such methods include ancestral recombination graphs (ARGs) advanced by Kuhner et al, 2000, and continued development of standards to represent ARGs e.g. McGill et al 2013 has contributed to further development of these methods (McGill, Walkup and Kuhner, 2013). Integration of recombination models and mechanistic epidemiological models remains challenging as it requires consideration of all factors that may have an impact on recombination.

1.7.3.5 RESOLVING WITHIN AND BETWEEN HOST EVOLUTION

The conventional application of principle of coalescence theory in fast evolving viruses from a phylodynamics modelling perspective has been the assumption that coalescent events in the virus phylogeny mirror the time of transmission events. This is somewhat true in cases where there is no significant genetic variation within the infected host. This is a challenge in phylodynamics given that in most viral infections, there is always significant genetic variation in the infected hosts hence caution needs to be taken when interpreting coalescent events as

transmission events (Volz, Koelle and Bedford, 2013; Ratmann *et al.*, 2017; Volz, Romero-Severson and Leitner, 2017a).

Distinguishing within and between host evolution has previously been hampered by the lack sequence data derived from both scenarios thus limiting the ability of investigators to resolve within and between host evolution (Frost *et al.*, 2015; Rife *et al.*, 2017; Volz, Romero-Severson and Leitner, 2017a). These types of challenges are set to be overcome in recent times due to availability of sequence data in both instances brought about by the speedy and decreasing cost of sequencing and therefore studies of pathogen evolution at multiple scales are now possible. Even with the availability of these abundant sequence data, combination of data at different scales remains a challenge to implement in some studies (Volz, Romero-Severson and Leitner, 2017b). This raises concerns of applying models to small well sampled populations in cases such as in outbreak situations.

The field of phylodynamics remains inadequately understood if the significant challenges of implementing models that capture different mechanisms for evolution at multiple scales because patterns of between host evolution do not just reflect a rescaling of within host patterns (Baele *et al.*, 2018).

Many factors need to considered and these may include: host-specific immune responses, temporal changes in selection pressure during the course of infection, founder effects, biased transmission of specific pathogen variants, and in the case of retroviruses such as HIV-1, the storage of the virus in the body. All these factors may have a significant impact on rate of within host-evolution compared to between-host evolution (Lemey, Rambaut and Pybus, 2006).

1.7.3.6 UPDATING ANALYTICAL TECHNIQUES IN TANDEM WITH TECHNOLOGICAL CHANGES IN SEQUENCING

There have been technological developments that have made it possible for large quantities of sequence data to be generated timeously, however the analytical strategies have lagged behind resulting in significant problems for phylogenetic inference. For example, there are large quantities of sequenced genomes for HIV and Influenza A viruses that have formed the basis for the development of phylodynamics analysis approaches.

A number of computational techniques have been implemented to improve the turn-around times of phylodynamic analyses such as the use of graphics processing units (GPUs) in computation tasks that considerably achieve reduction in the time taken to fit phylodynamic models (Ayres *et al.*, 2019). These models take considerably longer to complete given large datasets e.g. computation of phylogenetics likelihoods as applied in Maximum likelihood and Bayesian models (Baele *et al.*, 2018). The use of GPU computing approaches proves very helpful in cases where large Markov rate matrices are used for calculating the codon substitution models and/or analysing alignments with many sites (as opposed to many taxa) (Rife *et al.*, 2017). Current computing approaches still lack robustness and efficiency when applied to large numbers of sequences and therefore they require further refining to achieve optimal efficiency (Baele *et al.*, 2017).

In summary, the above phylodynamic challenges presents many opportunities for further model and analysis strategies improvement. As more model flavours become available or customized to address specific questions using a given datasets, many opportunities will be opened up for the analysis of 'big data' (Baele *et al.*, 2018). An impactful example would involve development of algorithms that can enhance proper mixing of Markov chain Monte Carlo approaches commonly used in estimation of the likelihood in Bayesian phylodynamic studies (Frost *et al.*, 2015).

CHAPTER 2

SPATIOTEMPORAL TRANSMISSION PATTERNS OF VIRAL INFECTIOUS DISEASES IN AFRICA WITH SPECIAL FOCUS ON THE INTRODUCTION AND DISPERSAL PATTERN OF THE 2009 INFLUENZA A/H1N1 PANDEMIC VIRUS IN AFRICA

2.1 INTRODUCTION

2.1.1 THE INFLUENZA A/H1N1 2009 PANDEMIC IN AFRICA

The first pandemic of this century was caused by 2009 influenza A/H1N1 virus which became to be commonly referred to as H1N1pdm strain of the H1N1 subtype (https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en). The Influenza A/H1N1 2009 pandemic virus was unique in many aspects with regards to its genetics and evolution (Trifonov *et al.*, 2009). The survival mechanism of Influenzaviruses is characterised by two main mechnisms: mutation and ressortment (Merler *et al.*, 2011; Liang *et al.*, 2013; Wille *et al.*, 2013). These mechanisms played a critical role in both the emrgence and spread of the 2009 A/H1N1pdm virus (Jombart *et al.*, 2009).

The genetic history of the pandemic H1N1 2009 was associated with reassortment events involving North American and Eurasian swine viruses. The implicated parental viruses of the pandemic H1N1 were also shown to have themselves arisen from reassortment events . Thus the present H1N1 pandemic viruses are known as triple reassortant strains (Morens, Taubenberger and Fauci, 2010). The pandemic (H1N1) 2009 (H1N1pdm) virus arose from a reassortment of two swine influenza viruses, namely, Eurasian H1N1 and a H1N2 circulating in North America, each of which themselves arose from reassortments (Balish *et al.*, 2009). Several studies have comprehensively reported different subtypes that may have contributed to the gene constellation of the 2009 H1N1 pandemic virus (Smith, Bahl, *et al.*, 2009; Smith, Vijaykrishna, *et al.*, 2009; Brett N Archer, Tempia, *et al.*, 2012; Su *et al.*, 2015b)

The swine H1N1 virus isolated from North America is predicted to have arisen from at least two prior reassortments in swine and comprehensive analyses have demonstrated that it contributed six segments: PB2, PB1, PA, HA, NP, and NS to the 2009 H1N1pdm virus (Garten *et al.*, 2009; Trifonov *et al.*, 2009). The known triple-reassortant swine H3N2 was first detected

in 1998; it originated from genome reassortment of classical swine H1N1 (contributing NS, NP and MP), avian H1N1 (PB2 and PA) and human H3N2 (HA, NA and PB1) (Garten *et al.*, 2009; Neumann, Noda and Kawaoka, 2009; Smith, Bahl, *et al.*, 2009; Trifonov, Khiabanian and Rabadan, 2009; Takebe *et al.*, 2010). Subsequently, the immediate North American swine progenitor of H1N1pdm, i.e., H1N2, was first detected in 1999 (Gibbs, Armstrong and Downie, 2009). It is a reassortant of triple reassortant H3N2 (PB2, PB1, PA, NP, NA, M and NS) and classical swine H1N1 (HA) (Smith, Bahl, *et al.*, 2009). The Eurasian swine H1N1 contributing to the pandemic H1N1 virus was also shown to have a mixed ancesstry arising from two reassortments, with both NA and MP being transferred from host avian, but at different times (Chen and Shih, 2009). In summary, it appears that the virus responsible, 2009 pandemic H1N1 (H1N1pdm), was the result of multiple reassortment events that brought together genomic segments from classical H1N1 swine influenza virus, human seasonal H3N2 influenza virus, North American avian influenza virus, and Eurasian avian-origin swine influenza viruses.

Although the 2009 influenza A/H1N1 pandemic virus emerged in Mexico and the USA in March and April 2009 respectively, it subsequently spread rapidly across the globe (Jombart *et al.*, 2009). By June 2009, laboratory confirmed cases had surpassed 30000 cases across 74 countries, and this prompted the WHO to declare it a pandemic (Christman *et al.*, 2011).

A global estimate of 49 discrete introductions of pH1N1 viruses from humans to swine from 2009 to 2012 is likely an underestimate, and increased surveillance efforts in swine continue to identify additional introductions of pH1N1 viruses from humans to swine globally (Nelson *et al.*, 2012, 2015).

In Africa, as with most other parts of the world, it was first detected between May and July 2009 (Barakat *et al.*, 2012; Brett N. Archer *et al.*, 2012; Katz *et al.*, 2012; Nzussouo *et al.*, 2012). A few countries in each African sub-region reported the first laboratory confirmed H1N1 pandemic virus incidences between June and July 2009 (Heraud et al., 2012). Several studies have reported regional and country specific detection and transmission timelines for this epidemic. Nzussouo et al 2015 reported a delayed onset of pandemic H1N1 2009 in the west African region from data analysed across 10 west African countries in which some countries recorded their first laboratory confirmed case of the pandemic virus 6 moths after its detection in Mexico (Nzussouo *et al.*, 2012). These results were echoed by a similar study in Morocco that demonstrated a similar delay in local transmission of pandemic A/H1N1 2009 virus during the epidemic and it further reports that A/H1N1 became the dominant strain during

2009-2010 Influenza season (Barakat *et al.*, 2012). Wong et al 2012 described the temporal and geographic progression of A(H1N1)pdm09 as it emerged in Kenya and characterized the outpatient population with A(H1N1)pdm09 infection and their study concluded that this virus was more prevalent in school going children than other epidemiological groups in contrast to the seasonal A/H1N1 and A/H3N2 circulating at that time (Wong *et al.*, 2012).

Although data on the emergence of this epidemic remain scanty in most African countries due to the lack of infrastructure needed to carry out rapid molecular diagnostic influenza tests; the sequence data available in public repositories from a few African countries provide an unprecedented opportunity to investigate the transmission dynamics of this virus in Africa. After the 2009 H1N1 outbreak, influenza surveillance in Africa has greatly improved (Radin, Katz, Tempia, Talla Nzussouo, *et al.*, 2012). Data from a sizeable number of African countries is now available and continued influenza sequencing activity within these countries provide evidence informed quantification of Influenza burden in Africa (Talla Nzussouo *et al.*, 2017).

Crucially, such studies may partially mitigate the generalized lack of influenza surveillance in Africa and help guide targeted pandemic intervention strategies. The reason for this is that viral genetic sequence data provides a valuable source of information on the historical record of epidemic influenza spread. The utility of such methods has already been demonstrated in several other studies investigating the molecular epidemiology of influenza and several other viral infectious diseases (Grenfell, 2004; Jombart *et al.*, 2009; Hedge, Rambaut and Lycett, 2013; Faria *et al.*, 2014).

The growing availability of sequence data and the continual improvement of phylogenetic techniques and models has provided new avenues for studying the molecular epidemiology of viral epidemics (Baillie *et al.*, 2012; Viboud *et al.*, 2013; du Plessis and Stadler, 2015). Using these approaches, the transmission dynamics of "measurably evolving" pathogens can be explored in detail at both the individual host and the population level. In an outbreak situation, the first goal usually involves identifying the spatial, temporal and genetic origins of the epidemic lineage. The HA-NA gene sequences play a crucial role in the transmission dynamics of influenza acros a wide range of species. Using HA sequences Su et al, 2015 reported unque genetic features of 2009 pandemic H1N1 illustrating its host jump ability and immune selection driven by nonsynonymous substitutions across the NA, M2 and NS genes (Su *et al.*, 2015a)

2.1.2 ECOLOGICAL, ECONOMIC AND GENETIC PREDICTORS OF TRANSMISSION PATTERNS OF OF 2009 INFLUENZA A/H1N1PDM IN AFRICA

The spread of infectious diseases is potentially influenced by multiple factors (Wu *et al.*, 2013), including environmental, genetic, demographic, economic or eographic variables. It has been demonstrated in several modelling studies that climate change, for instance, has had an impact on the emergence and re-emergence of infectious disease outbreaks (Wu *et al.*, 2013; Gebreyes *et al.*, 2014; Li, Grassly and Fraser, 2014). Another example is the transmission dynamics of rabies in North Africa which was shown to be influenced by human movements with transmission bottlenecks and regional borders impeding of spatial transmission of pathogens (Talbi *et al.*, 2010).

Determining the relative contributions of environmental, genetic, demographic, economic or eographic variables to disease spread is challenging even in cases where all the potentially contributing factors are known, but is considerably more difficult if some of these have not even been identified. Nevertheless unravelling the factors that have high propensity to contribute to the spread of infectious diseases remains a priority with respect to informing disease outbreak mitigation and management strategies.

To meet this need, bayesian statistical models have been devised and computationally implemented that aim to simultaneously test the contribution of any number of potentially disease modulating factors on the transmission history or observed patterns of spread of viral infectious diseases (Lemey *et al.*, 2012; Beard *et al.*, 2014). These "generalized linear models" (GLMs) are an improvement on simpler phylogeographic models that were devised to just explain the spatial dispersal of pathogen using a combination of sequence and sampling location data (Lemey *et al.*, 2012). Although earlier spatial epidemiology models have enabled researchers to obtain insights into the impact of factors such as human mobility or environmental conditions on disease spread, they did not use sequence data and therefore failed to access and use information encoded in this genetic data to test the relative contributions of different predictor variables (Beard *et al.*, 2014; Magee *et al.*, 2014). GLMs enable spatial epidemiological information to be analysed simultaneously with sequence data and yield information on the underlying genetic (host or pathogen), socio-economic, climatic, ecological, or geographical causes of disease outbreaks.

Specifically GLMSs enable assessment of the association between individual predictors and the evolution and/or spatial diffusion of viruses. Although GLMs cannot be used to determine

causality *per se*, the nevertheless can specifically identify the likely drivers of outbreaks and, on this basis, the insights they potentially yield could be extremely valueble for formulating interventions to prevent, contain and monitor disease outbreaks (Lemey *et al.*, 2014).

GLMs works by allowing the parameters of an instantaneous rate matrix to be transformed as logarithms of a combination of a set of selected epidemiological, ecological or economic predictors (Lemey *et al.*, 2012, 2014; Beard *et al.*, 2014; Nunes *et al.*, 2014). During evaluation of the contribution and weight of each predictor for the dispersal process, two statistics, the inclusion probability and the conditional effect size, are calculated. The support for each predictor is estimated by comparing the prior values of these statistics with their posterior values using a Bayes Factor test (Lemey *et al.*, 2012).

So far these GLMs have been productively used to study the transmission dynamics of a number of different viral pathogens. They have consistently revealed that increased degrees of human mobility is a major factor driving the transition from localised outbreaks widespread epidemics (Nunes et al., 2014; Philippe Lemey et al., 2014; 2012). For example, air travel in particular has played a significantly more important role in the global dissemination of seasonal Influenza A/H3N2 viruses and in the 2009 A/H1N1 global pandemic, than factors such as average or minimum distance between locations and population size or density which only have strong impacts on the transmission of the virus at local scales (Lemey *et al.*, 2012, 2014). GLMs also revealed that the most important contributors to the circulation of Influenza A/H1N1 viruses in Egypt were a genetic motif near the cleavage site of the HA gene and human/avian host densities (Magee *et al.*, 2014). Other associations revealed using GLMs include the identification of Porto Allegre in Brazil as the source of the HIV-1C epidemic in that country (Gräf, Vrancken, Junqueira, *et al.*, 2015); that geographical distance from areas of active transmission and attenuation accruing from restricted human movement at international borders were the two key factors containing the 2014-2015 Ebola brus outbreak in West Africa (Dudas *et al.*, 2017).

On the African scale, several factors may influence the spread of infectious diseases. These may range from ecological, economic and demographic factors. Aerosol tranmistted infections such as influenza may be heavily impacted by climatic conditions especially in temperate regions which experience clear seasonal patterns. In tropical regions these seasonality is less apparent and a study by Monamele et al 2017 showed very weak association influenza activity with meteorological variables such as humidity, rainfall and temperature suggesting that

influenza in tropical zones may occur throught the year irrespective of the season (Monamele *et al.*, 2017). A smilar study conducted in northern Cameroon testing the influence of three meteorological factors ie temperature, rainfall and relative humidity reported a significant association between relative humidity and Influenza A activity but no correlation was observed between the same variables and inicidence of Influenza B activities (Munshili Njifon *et al.*, 2018). In a recent Ebola Virus Disease (EVD) outbreak in west Africa, Valeri et al 2016, tested the association of estimated epidemic parameters using mathemat-ical models, and estimated their associations using ecological regression models to predict Subnational Ebola Epidemic Dynamics. In this study, they identified some factors predicting rapid and severe EVD epidemics in West African subnational regions. All the tested factors did not show significant contribution to the spread of EVD during the outbreak however using stepwise multivariable models only mean education levels were consistently associated with a worse local epidemic (Valeri *et al.*, 2016). An additional study that applied statistical epidemic epidemic modelling, investigated population level predictors of EVD risk at the regional level in Sierra Leone, Liberia, and Guinea (Levy and Odoi, 2018). In this particular study, spatial and descriptive analyses were conducted to assess distribution of EVD cases and their findings suggested that the risk of EVD was significantly lower in areas with higher proportions of: (a) the population living in urban areas, (b) households with a low quality or no toilets, and (c) married men working in blue collar jobs. However, one surprising finding was that the risk of EVD was significantly higher in areas with high mean years of education thus contrasting the findings of sister study aforementioned above. In the current study, I set out to test a combination of geographic, genetic, economic and demographic factors that may have impacted the observed spread of the 2009 Influenza A/H1N1 pandemic virus in on continent wide scale thus expanding the attempts by previous studies whose focus was on subnational national or regional scale.Here, I make use of sequence data from three membrane-associated Influenza A virus genes that play a crucial role in driving Influenza A virus infections: hemagglutinin (HA), neuraminidase (NA) and matrix protein (M1/M2). My two specific intentions were: (1) to gain insights into the introduction and the dissemination pattern of influenza A H1N1 within Africa; and (2) to identify key demographic, ecological, economic and social factors that influenced the transmission pattern of this virus within Africa following the start of the 2009 pandemic.

2.2 MATERIALS AND METHODS

2.2.1 SEQUENCE RETRIEVAL, SELECTION AND ALIGNMENT

An online search of three public sequence repositories was conducted using the search terms '2009 H1N1 pandemic' or 'swine H1N109' in order to access the Influenza A/H1N1 pandemic sequences from Africa. These included NCBI's Influenza virus resource (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?cmd=database), Influenza research database (IRD) accessible at (www.fludb.org) and the global initiative on sharing all Influenza data (https://www.gisaid.org/). The GISAID database contained the most comprehensive sequence datasets relating to the 2009 H1N1 pandemic from Africa. I retrieved all the sequences submitted to the GISAID database during the first and second waves of the pandemic on the 23rd July 2014. Additionally, I also retrieved post-pandemic sequences. This search returned 760 hemagglutinin (HA), 514 neuraminidase (NA), and 300 matrix protein (MP) sequences of African origin. Other segments returned fewer sequences that were generally not sampled evenly throughout the entire African region and these were therefore not used in further analyses. Using the same approach, we retrieved sequences for the HA, NA and MP encoding segments from other non-African regions around the world that were collected and submitted to the database over the same period. There were more than 4000 non-African sequences available for each of these segments and given that the intended Bayesian analysis is memory-intensive, it was therefore necessary to subsample to obtain a subset of these sequences for further downstream analyses. I therefore performed selection of sequences from phylogenies reconstructed using Fasttree (Price, Dehal and Arkin, 2010) after combining all the African and non-African sequences for each of the three segments aligned using MUSCLE (R C Edgar, 2004) . From these phylogenetic trees, I identified and selected all the non-African sequences that clustered together with African sequences within these trees. This yielded 222 HA, 319 NA and 120, MP sequences from non-African regions (including Asia, Europe, North America, Oceania and South America) which were used as representatives of the globally circulating H1N1 viruses. The final datasets used for subsequent analyses comprised 982 HA, 833 NA and 420 MP sequences. These were separately aligned using MUSCLE (Robert C Edgar, 2004) and visualized in Seaview (Gouy, Guindon and Gascuel, 2010) before undertaking further analyses.

2.2.2 PHYLOGENETIC AND PHYLOGEOGRAPHIC ANALYSIS

2.2.2.1 EVOLUTIONARY ANALYSIS

To gain insights into the evolutionary dynamics of the 2009 Influenza A/H1N1 pandemic viruses circulating in Africa during the epidemic and post epidemic periods, I conducted temporal evolutionary analysis on the HA, NA and MP alignments that I had produced. The dates of sampling for each of the sequences in the alignment were used to generate time-scaled phylogenies as implemented in the computer program, BEASTv1.8.2 available at {http://tree.bio.ed.ac.uk/software/beast/)(Drummond *et al.*, 2012). The genetic origin of the viral lineages circulating in Africa were determined by applying a GTR + G + I nucleotide substitution model (selected as the best fit model by Jmodeltest (Posada, 2008)), an uncorrelated relaxed molecular clock model (Ho *et al.*, 2011) and a coalescent Bayesian skyline demographic model for modelling virus population changes over the sampling period (Drummond, Rambaut and Xie, 2011; HO and SHAPIRO, 2011). BEAST analyses were carried out by five independent runs with each run lasting for 500 million generations with every 50000th tree being sampled for further analysis and annotation. As this involved a Markov chain Monte Carlo (MCMC) based approach, convergence and proper mixing of the chain was essential for reliable phylogenetic and evolutionary inferences. I assessed mixing and convergence using Tracer version 1.6 (Rambaut A, Suchard MA, Xie D & Drummond AJ, 2014) which is available from (http://beast.bio.ed.ac.uk/tracer). Once convergence and good mixing was achieved (as inferred by effective sample sizes (ESSs) greater than 200) for each of the five independent runs, these were combined (this is acceptable since they had converged on approximately the same likelihood) after discarding ~30% as burn-in to achieve ESSs >200 for all the parameters that were being estimated by the models being applied. The combined trees were summarised using the computer program TreeAnnotator (which is part of the BEAST package) available at (http://beast.bio.ed.ac.uk/TreeAnnotator) to generate a phylogeny commonly referred to as maximum clade credibility (MCC) tree. The MCC tree was visualised in Figtree available at (http://tree.bio.ed.ac.uk/software/figtree/) and further annotations were made with the aim of describing the observed evolutionary and temporal aspects of the datasets under study.

2.2.2.2 SPATIAL ORIGINS ANALYSIS

Understanding the geographical source of the circulating genotypes of a pathogen during an outbreak such as the 2009 Influenza A/H1N1 pandemic remains an important approach in characterizing, managing and controlling the outbreak resulting from such pathogens. Sequence data collated as

described in section 1.2.0 were mapped to their respective sampling locations (regions). The African region was further divided into five sub-regions: central, eastern, northern southern and western regions. The regions following geopolitical demarcations with the goal of maximizing the spatial resolution of virus circulation in Africa. The non-African sequences were also tagged according to the continental region from which they were isolated: Asia, Europe, North America, Oceania and South America. Spatial diffusion of viral pathogens within locations can be simulated by a continuous-time Markov chain (CTMC) process along each branch of the viral phylogeny. Using these locations as discretized states, we performed a discrete phylogeography analysis that utilized a Bayesian Stochastic Search Variable Selection (BSSVS) statistical approach to test random transitions (movements) between pairs of discrete states (i.e. locations in this case) by allowing some transitions to be zero with some weighted probability i.e. turning on or off of some indicators during the modelling process so as to allow the assessment and quantification of confidence in the form of Bayes factors for each transition pair in the location matrix (P Lemey *et al.*, 2009). This was implemented in BEAST that has the capacity to integrate multiple types of dataset under the same statistical framework. I assumed symmetrical movement between any pair of sampling locations (which is plausible given the mobility of human populations). Since we were interested in making inferences of the origins and dissemination pathways of this virus into Africa and out of the African region, the option to infer the ancestral location at every node was chosen rather than the root location states alone. The spatial patterns were further analyzed in SPREAD (Bielejec *et al.*, 2011) to identify any statistically supported movements between any pairs of locations. SPREAD was also used to generate Keyhole Markup Language (KML) files that could be used to visualize dispersal pathways of the virus among sampled locations on geodetic programs such as Google Earth (Conroy *et al.*, 2008).

2.2.3 Analysis of the Diffusion patterns of H1N1 in Africa

In an infectious disease outbreak situation, some of the most important pieces of information necessary for guiding timely interventions are the dispersal rate, directionality, and the distances covered by the infectious disease agent. When sampling locations for a particular pathogen are known, modelling the diffusion between the locations can only be achieved using Markov processes if the observed locations are treated as continuous states (Lemey *et al.*, 2010).

I aimed at uncovering the rate of dispersal of the H1N1 pandemic virus across the African continent by reconstructing the spatial diffusion in continuous space and time during the peak epidemic and post-epidemic periods i.e. 2009-2014. Spatial reconstruction in continuous space and time has the added advantages (1) of revealing the pathogen's diffusion process at any particular point in time and (2) that high probability regions for ancestral locations at arbitrary

times in the diffusion process can be identified i.e. it alleviates the limits imposed by treating sampling locations as discrete states and allows inferences to be made about the dispersal process even in the unobserved locations (Lemey *et al.*, 2010; Nunes *et al.*, 2014). I conducted spatial reconstruction analyses using both the uniform diffusion process model-Brownian Diffusion (BD) and alternative models that relax the assumption of uniform dispersal processes and allow for variations in the dispersal process under a variety of distribution models e.g. Cauchy, gamma, and lognormal models popularly known as relaxed random walks (RRWs) (Lemey *et al.*, 2010). Each sequence in the in the three alignments (HA, MP and NA) was coupled to its global positioning system (GPS) coordinates. We performed four independent continuous phylogeographic analyses in BEAST version 1.8.2 under the GTR + G + I nucleotide substitution model, nonparametric coalescent Bayesian skyline demographic model and either homogenous or Cauchy, gamma or lognormal models for modelling of the spatial diffusion patterns. These were run for 500 million generations with parameters and trees being sampled every 50000 generations. To identify the best model supporting the spatial diffusion process in continuous space for this virus in Africa, we used path sampling (PS) and stepping stone sampling (SS) procedures as described in (Baele *et al.*, 2012). We assessed convergence and proper mixing in Tracer version 1.6 and generated annotated MCC trees in TreeAnnotator v1.8.2 and visualised these trees in Figtree v1.4.2. The spatial reconstructions were further analysed in SPREAD and the generated KML files were visualised in Google Earth.

2.2.4 Investigation of the predictors of H1N1 dispersal in Africa

The African continent is faced with a myriad of challenges when trying to tackle an outbreak on the scale of the 2009 Influenza A/H1N1 epidemic. Among others, ecological, economic, and epidemiological factors are known to play a role in the spread of pathogens and determining the relative contributions of each of these could inform future targeted interventions using whatever limited resources are available. Testing the relative contributions of various disease transmission factors can be achieved with the use of generalised linear models (GLM). As it is implemented in BEAST 1.8.2 the GLM algorithm is able to integrate various factors into a Bayesian phylogenetic framework where the instantaneous rate matrix is formed by taking the logarithms of a set of factors/predictors which are thought might be involved in the diffusion of a particular pathogen (Beard *et al.*, 2014; Lemey *et al.*, 2014; Magee *et al.*, 2014; Nunes *et al.*, 2014) (see equation below) .

$$\log L_{ij} = b_1 \P_1 x_{i,j,1} + b_2 \P_2 x_{i,j,2} + \ldots + b_p \P_p x_{i,j,p}$$

where:

L = *rate matrix of discrete location change* (K x K)

$b = (b_1, \ldots, b_p)'$ = *effective sizes*

$\P_1, \ldots, \P_p$ = *indicator* $(0,1)$

where K is the number of discrete location states under consideration.

This is done while simultaneously reconstructing the evolutionary history and averaging over uncertainty at both the phylogenetic and dispersal pattern levels. The contribution of each factor or predictor to the diffusion pattern can then be quantified by assigning an inclusion probability and conditional effective size to each predictor. The weight of each factor was obtained by calculating Bayes Factors of the prior in comparison with the posterior expectation or inclusion probability.

2.2.4.1 POTENTIAL ECOLOGICAL, ECONOMIC AND GENETIC PREDICTORS TESTED

Data on a total of 16 predictors that could have plausibly contributed to the dispersal of the 2009 Influenza A/H1N1pandemic virus amongst the African countries where these viruses were sampled was compiled from public databases. These predictors included:

1. **Geographical factors** e.g. Agglomeration index, location (latitude/longitude separately), great circle distances, total population sizes and road distances;

2. **Economic factors** e.g. Gross domestic product (GDP), trade (exports and imports);

3. Infrastructural factors e.g. Number of cars per 100km of road network, number of departing flights per year, number of air passengers per year, railway coverage;

4.**Genetic factors** e.g. Number of sequences (sample sizes), flu vaccine coverage and incidence (Number of H1N1 pandemic Influenza cases detected).

Agglomeration Index

Infectious disease spread is influenced by the mobility of infected hosts. The movement of these hosts may create an environment in which there can be transmission of diseases between infected and susceptible hosts. The agglomeration index is a measure of population concentration in specific geographic areas of a country and is calculated based on total

population size in urban centres, population density and travel time to urban centres. Urban areas are considered to have higher agglomeration indices as many households occupy small spatial areas as opposed to rural areas in which homesteads, schools, churches or markets are many kilometres apart. I considered the average agglomeration index of the 26 African countries from which influenza virus sequences were sampled. Country-specific agglomeration indices data were retrieved from the World Bank World Development Indicator report available at http://siteresources.worldbank.org/INTWDR2009/Resources/4231006-1204741572978/Hiro1.pdf (Uchida and Nelson, 2008).

Latitude and Longitude

Geographical coordinates of the centroid of each sampled country were obtained using the Google Maps interface implemented in an in-house python script. Although the exact sampling locations did not coincide with centroids, centroids were used to achieve uniformity in the data integrated into the GLM models.

Great circle distances (as the crow flies)

Pairwise geographical distances (in kilometres) were estimated using Google maps considering the curvature of the Earth (as opposed to Euclidean distances which would have implied movement distances that pierced the earth's surface.

Total population

The total population size for each the sampled countries was obtained from the World Bank database. It is postulated that, relative to less-populous countries, highly populated countries have a higher proportion of individuals that are at risk of contracting and spreading an infectious disease during outbreaks. To explore this, total population sizes of the sampled countries were included as a predictor in the GLM model.

Road distances

Google Maps were used to calculate the pairwise driving distance between the centroids of all the sampled countries. Although direct road connection in some pairs of countries are non-existent, we nevertheless estimated the road travel distance to the nearest capital city of that country. For island states such as Madagascar, Mauritius and the Seychelles a lack of road connectivity was handled with the use of large distances values of one million kilometres.

Gross domestic product

The gross domestic product of a country is a metric that reveals the country's economic output. This metric was considered for testing in the GLM model as a possible predictor of infectious disease transmission in recognition of the fact that resources in countries with higher GDP could be channelled to mitigate spread of infectious diseases during outbreaks more swiftly than could those of countries which more limited resources. Towards this end, the GDP of each sampled country were retrieved from the World Development Indicator (WDI) database (http://databank.worldbank.org).

Trade-exports and imports

Trade facilitates the movements of people and goods across borders and this could play a role in the dissemination of infectious agents if infected people or animals are involved in cross-border movements. Intra-Africa trade statistics for each of the sampled countries were obtained from the United Nations Conference on Trade and Development (UNCTAD) database (www.unctad.org)

Number of cars

I obtained data on the total number of cars in each of the sampled countries as a proxy for the number of people using road transport within the individual countries. This metric is also an indirect measure of economic development and the influence of road transport on the dissemination of infectious disease agents such as the one currently under study.

Number of departing flights

Air transport has been shown to play a significant role in the diffusion of viral infectious diseases both globally such as with Influenza A/H3N2; (Lemey *et al.*, 2014), and on a regional scale such as with Dengue virus; (Nunes *et al.*, 2014). In this regard, I opted to test the number of departing flights as a predictor of disease spread. Flight data was retrieved from the World Bank database's mode of transport statistics covering both the A/H1N1 pandemic and post-pandemic periods on yearly basis.

Number of air passengers

The number of air passengers represents the movement of either infected or susceptible hosts. Rapid and frequent travel to and from outbreak areas both provides opportunities for infected individuals to transmit pathogens over larger areas and exposes susceptible individuals to greater risk of acquiring infections. Data on annual volumes of passengers traversing through

each of the African countries sampled in this study was obtained from World Bank database's transport statistics (http://databank.worldbank.org) and were tested as a predictor of disease transmission patterns.

Railway coverage (km)

Railway transport is still a major form of travel in many African countries. Data on railway coverage was used as proxy for the utilisation of railway transport. For each of the sampled countries, railway coverage (in kilometres) was retrieved from World Bank Development database (http://databank.worldbank.org).

Number of sequences

To test the influence of sampling biases on the inference of the most probable predictors of observed spread, the number of genetic sequence samples for each location (i.e. country) was included as a predictor in the GLM model.

Influenza vaccine coverage

Mass vaccination has been known to be an effective intervention for preventing or slowing the spread of infectious diseases (Grabenstein and Nevin, 2006; Sander *et al.*, 2010; WHO, 2014) and therefore could have played a role in the pattern of Influenza A/H1N1 spread in Africa. Vaccines against the A/H1N1 strain became available in some African countries towards the end of first wave of the pandemic in November – December 2009. Vaccine coverage was calculated based on the vaccine doses supplied and used (data available in the WHO database) for individual African countries.

Number of Influenza cases (incidence)

From an infectious disease epidemiology perspective, the number of infected hosts generally plays a major role in the spread of pathogen. Therefore, incidence was tested as a predictor of viral movements. Numbers of confirmed H1N1 cases was obtained from the WHO flunet database (give the URL). Cumulative weekly incidence data of the 2009 Influenza A/H1N1 pandemic virus was accessed at http://www.who.int/influenza/gisrs laboratory/flunet/en/.

2.2.4.2 EVALUATION OF PREDICTOR INCLUSION

The GLM model analysis of potential predictors of Influenza A H1N1dissemination patterns was conducted using BEASTv1.8.2 and were based on posterior inclusion probabilities for each individual predictor included in the analysis. Bayes Factors were then calculated from

these posterior probabilities and were used to quantify the support of each predictor included in the GLM model. The inclusion probability is a statistical measure of the frequency with which an individual predictor was included in the model and was considered as a raw support statistic. The higher the inclusion probability, the more likely it is that the predictor is contributing to the observed diffusion process. Additionally, another quantity, denoted as **β**-coefficient, is also generated by the GLM analysis for each tested predictor and signifies the contribution of the predictor to the model when it is included. Each predictor in the model is assumed to have equal prior probability of being included and is referred to as binormial prior. The baseline binomial prior assigns a 50% chance that no predictor is included in the model (i.e. there is a corresponding 50% probability that it will be included in the model). The cut-off for a significant Bayes Factor was 10, as applied in other analyses employing BEAST e.g. (P Lemey *et al.*, 2009; Lemey *et al.*, 2010).

2.3 RESULTS

2.3.1 DATASETS

Nucleotide sequences representing the three membrane associated protein encoding segments of Influenza A H1N1 isolates (i.e. HA. MP and NA) were obtained from the public GISAID and NCBI IVR databases. Each sequence retrieved had to meet the following inclusion criteria: (i) it needed to span more than two thirds of entire segment; (ii) its date of collection needed to be available (resolved to at least the year and month in which it was sampled); and (iii) The geographical location where it was sampled needed to be available (resolved to at least its country of origin). A total of 760 HA, 320 MP and 513 NA, sequences were available from 26, 18 and 24 African countries respectively (Table 1). African countries were assigned to five African sub-regions: central Africa, eastern Africa, northern Africa, southern Africa and western Africa (Table 1, Table 2). The total number of sequences used to prepare alignments used for subsequent analyses were: HA (n=982), MP (n=420) and NA (n=833) (Table 2). The eastern African region had the highest number of sequences followed in order by the western, southern, northern and central Africa regions. African sequences were subsequently combined with closely related sequences sampled elsewhere in the world. Sequences from these other global regions were grouped according to their regions of isolation: i.e. Asia, Europe, North America, Oceania and South America. The rationale for this African vs non-African spatial demarcation was to achieve a higher intra-Africa spatial resolution while at the same time enabling the determination, on a continental-scale, of the locations from which different H1N1 lineages entered Africa.

Table 2.1: 2009 Influenza H1N1pdm HA, MP, and NA sequences collected in various African countries.

Subregion	Country	HA	NA	MP
Central Africa	Cameroon	41	7	39
Eastern Africa	Djibouti	2		
	Ethiopia	18	18	17
	Kenya	190	128	127
	Seychelles	1	1	1
	Tanzania	34	22	23
	Uganda	19	9	9
Northern Africa	Algeria	23	1	10
	Egypt	26	3	12
	Morocco	27	20	25
	Tunisia	49	1	16
Southern Africa	Angola	12		
	Madagascar	44	7	27
	Mauritius	16		10
	South Africa	78	20	36
	Zambia	12		
Western Africa	Burkina Faso	11	11	11
	Cote deIvoire	16	7	15
	Gambia	9		9
	Ghana	68	18	63
	Mali	1	1	1
	Mauritania	6		6
	Niger	4		4
	Nigeria	11	11	11
	Togo	3		4
	Senegal	50	15	37
Total		**760**	**300**	**513**

Table 2.2: Total number of sequences of the 2009 Influenza H1N1pdm used in these study

Location	HA	NA	MP
Central Africa	41	39	7
Eastern Africa	264	178	178
Northern Africa	118	67	25
Southern Africa	161	73	27
Western Africa	176	157	63
Asia	68	40	51
Europe	51	45	32
North America	87	200	20
Oceania	12	25	13
South America	4	9	4
Total	**982**	**833**	**420**

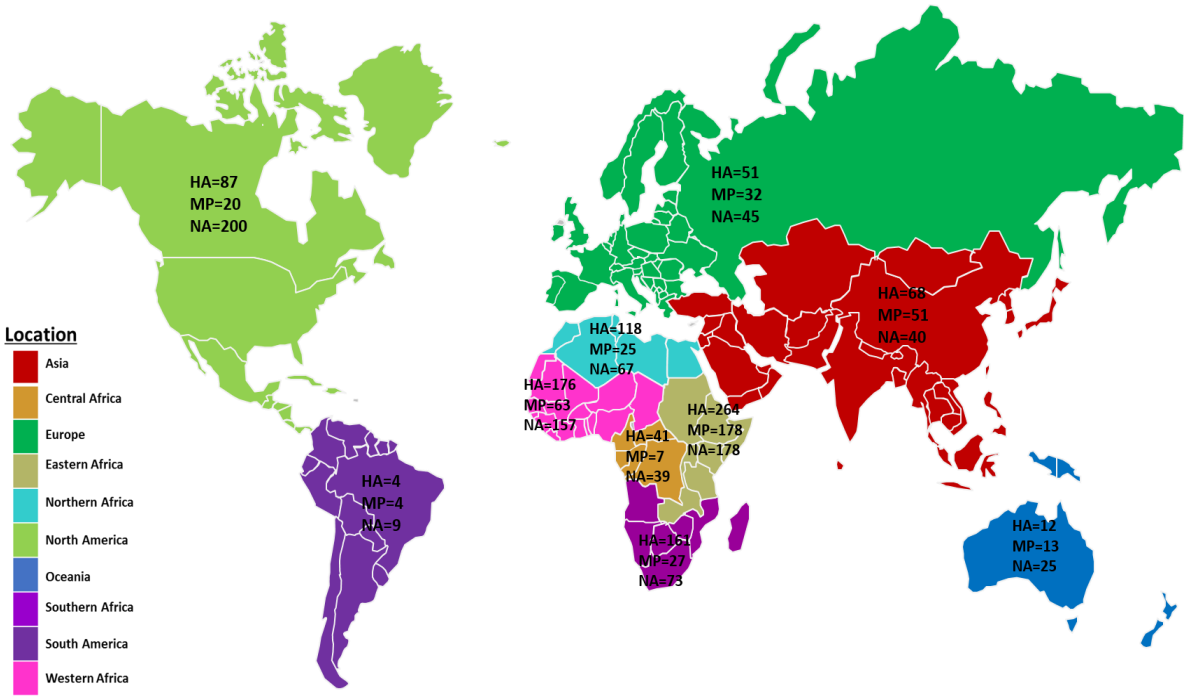


Figure 2.1: Geographical sampling locations

All sequences included in the analysis were assigned to one of ten locations; those sampled in Africa to the central, eastern, northern, southern or western African sub-regions and those sampled elsewhere to the Asian, European, North American, Oceanian or South American regions. The number of Hemagglutinin (HA), matrix protein (MP) and Neuraminidase (NA) sequences are shown for each region.

2.3.2 Evolutionary and temporal dynamics of 2009 Influenza A/H1N1 pandemic virus in Africa

2.3.2.1 Time scale of H1N1 pandemic virus introduction to Africa regions

Evolutionary timescales can be estimated from sequence data through molecular dating analyses. Patterns of evolutionary rate variation across lineages can be described by applying molecular clock models thus enabling relative ages of nodes in the phylogeny to be estimated. To this end, I explored the evolutionary and temporal dynamics of the 2009 Influenza A/H1N1 epidemic using these full or near full length HA, NA and MP sequences.

The mean evolutionary rates of the circulating 2009 Influenza A/H1N1 virus during the period of 2009-2014 (i.e. during both the pandemic and post-pandemic periods) were 6.039(±0.531) x 10^{-3} substitutions per site per year (subs/site/year) for HA, 4.652 (±0.912) x 10^{-3}subs/site/year for MP and 4.229 (±0.568) x10^{-3} subs/site/year for NA. The overall time to the most recent

common ancestors of the HA, NA and MP sequences included in these analyses were mid 2007, mid 2008 and early 2008 respectively.

A time-scaled phylogeny was reconstructed using the Bayesian phylogenetics approach implemented in BEAST which takes dated and geo-tagged sequence data as input under a GTR model of nucleotide substitution and assuming a flexible uncorrelated lognormal relaxed clock model to infer the rate of nucleotide substitution. A total of nine distinct clusters that reflect different circulating lineages were observed using the reconstructed HA phylogeny (Figure 2), similarly, five clusters from the NA phylogeny (Figure 3), and eight distinct clades from the MP phylogeny (Figure 4). These phylogenies revealed a high degree of H1N1 genetic diversity during the early stages of the epidemic between 2008 and 2010. Further inspection of the phylogenies suggests that after the year 2010, fewer lineages survived to maintain transmission and circulation in subsequent years. The earliest point of entry of H1N1 into Africa, as inferred from the reconstructed HA sequences, was northern Africa in approximately in April 2009 (posterior 95% credibility interval between 03/09 and 06/09) (table 3). The inferred median period of first introduction into the central African region was March 2010 (BCI 12/09-04/10) whereas the estimated time for introduction in northern, southern, western and eastern regions was estimated to be , April 2009 (BCI: 03/09-06/09), August 2009 (BCI: 05/09-10/09) and October 2009(BCI: 08/09-11/09) and November 2009 (BCI: 10/09-12/09) (Table 3).

Using MP sequences, the time-scaled reconstructed phylogeny of the 2009 Influenza A/H1N1 pandemic virus indicated introductions into the central, eastern, northern, southern and western African regions to have occurred in 02/11 (BCI 10/10 - 04/11), 09/08 (BCI 01/08 - 02/09), 10/09 (BCI 08/09 -11/09), 06/10 (BCI 04/10 - 07/10) and 08/09 (BCI 05/09 -10/09) (Table 3). Likewise, using the NA sequences, the inferred dates of introduction of H1N1 pandemic virus into the African regions: central, eastern, northern, southern and western regions were 01/10 (BCI: 11/09 -01/10), 09/09 (BCI: 06/10 -10/09), 10/10 (BCI 09/10 - 10/09), 09/09(BCI 08/09 - 09/09) and 10/09 (BCI: 09/09 -11/09) (Table 3).

Table 2.3: Inferred time of emergence (in month/year) of the Influenza A/H1N1 HA, MP and NA segments first entered various geographical regions

Clade	HA MRCA date (mm/yy)			MP MRCA date (mm/yy)			NA MRCA date (mm/yy)		
	MH	LHPD	UHPD	MH	LHPD	UHPD	MH	LHPD	UHPD
Central Africa	03/10	04/10	12/09	02/11	04/11	10/10	01/10	01/10	11/09
Eastern Africa	11/09	12/09	10/09	09/08	02/09	01/08	09/09	10/09	06/09
North Africa	04/09	06/09	03/09	10/09	11/09	08/09	10/09	10/09	09/09
Southern Africa	04/09	05/09	02/09	06/10	07/10	04/10	09/09	09/09	08/09
Western Africa	10/09	11/09	08/09	08/09	10/09	05/09	10/09	11/09	09/09

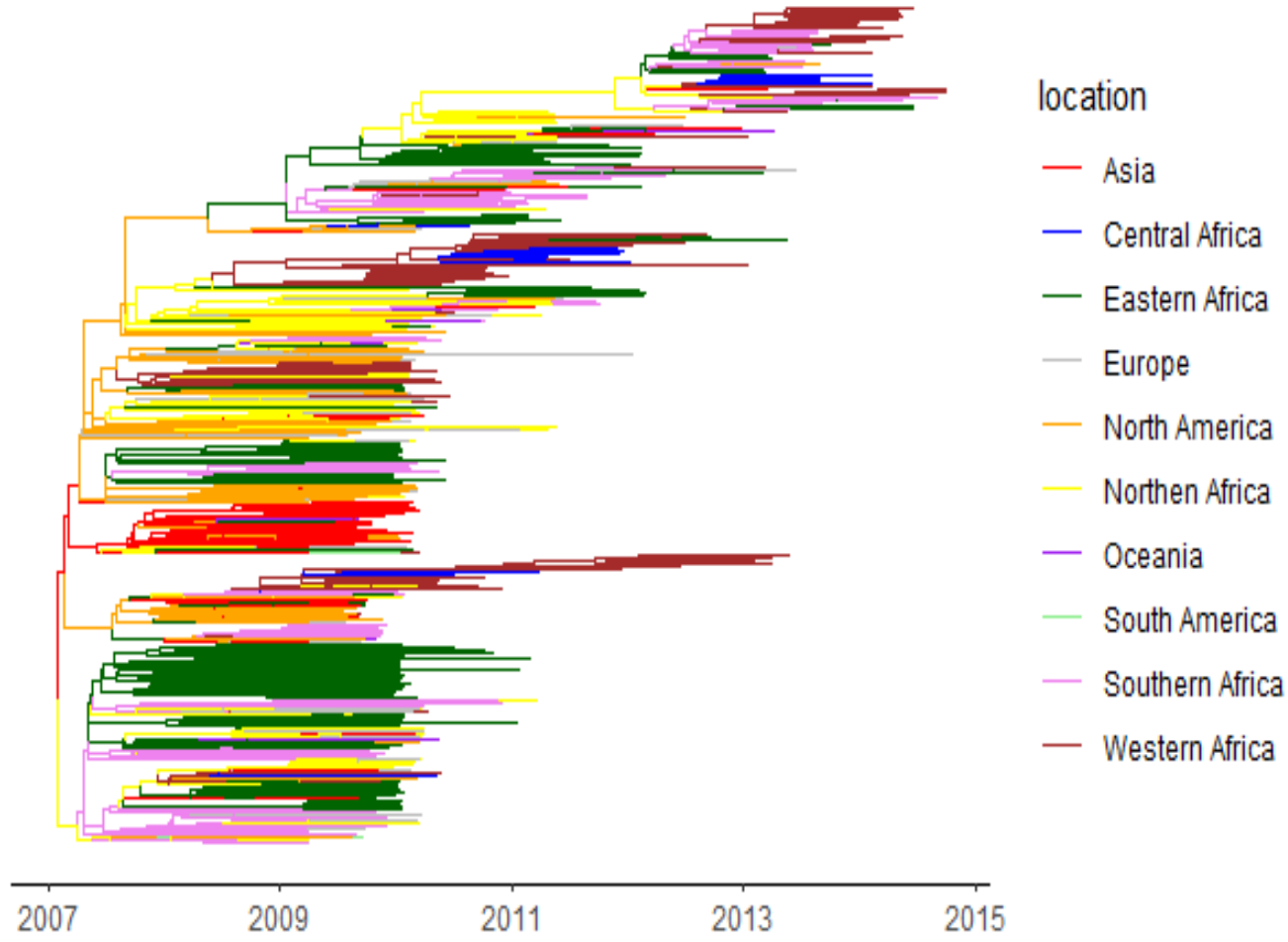


Figure 2.2: Bayesian Maximum clade credibility phylogeny of HA sequences generated under symmetrical discrete phylogeographic model.

Clades are coloured according to sample location. The geographical origin of the most recent common ancestor of this segment is inferred to be Asia. A total of nine distinct lineages were determined from the clustering pattern of the sequences in the reconstructed phylogeny.

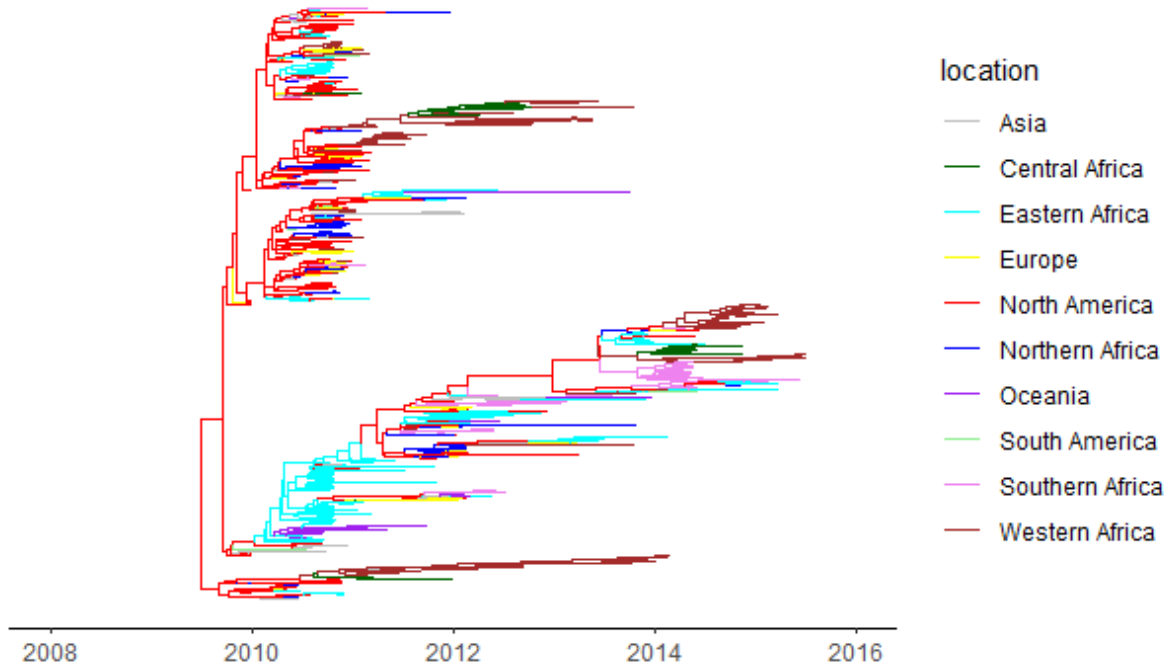


Figure 2.3: Bayesian Maximum clade credibility phylogeny of NA sequences generated under symmetrical discrete phylogeographic model.

Clades are coloured according to sample location. The MRCA of this segment is inferred to have existed in North America and the estimated tMRCA, suggest that virus had been in circulation a few years before its detection and it subsequent naming as the 2009 Influenza A/H1N1 pandemic virus in Mexico in April 2009.

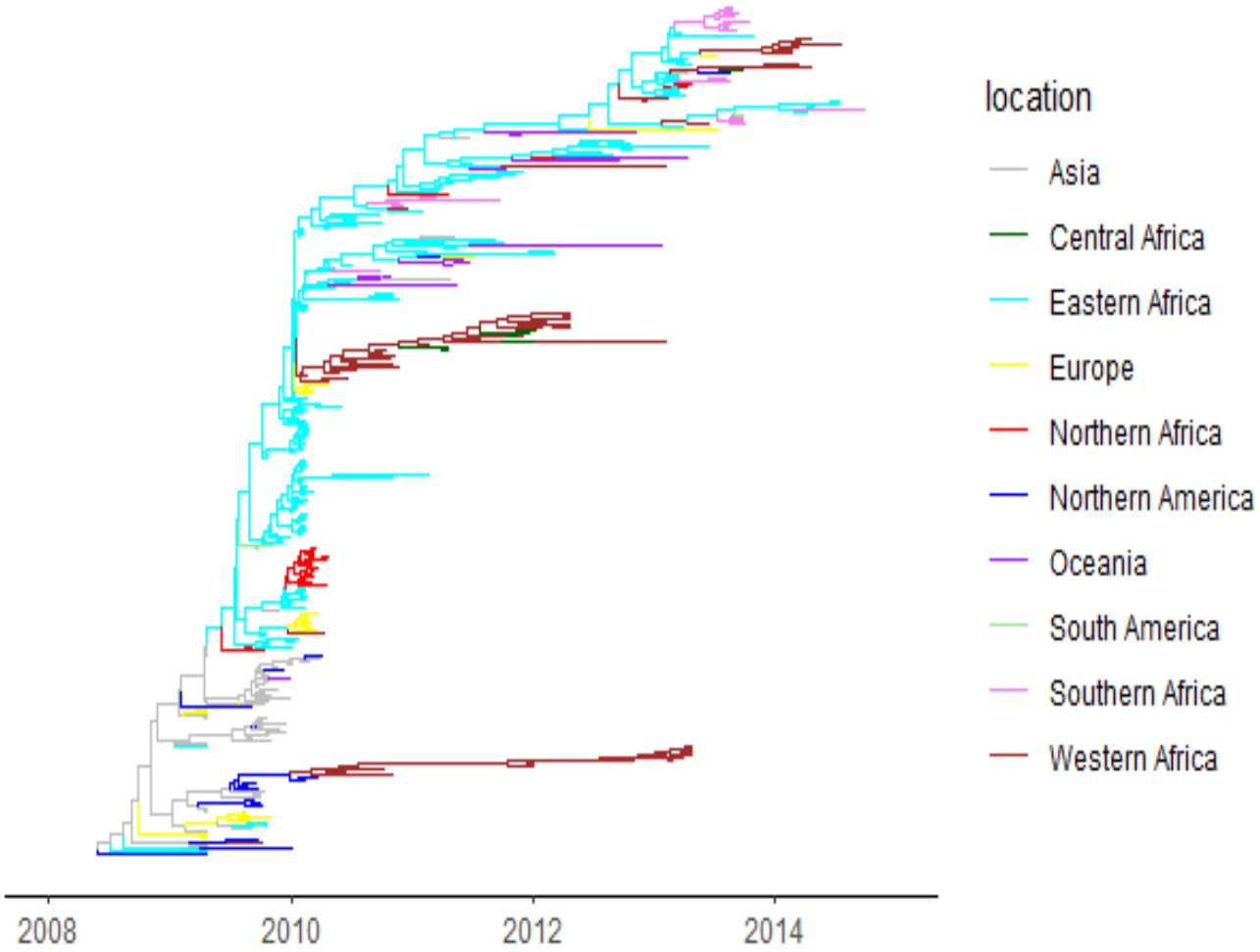


Figure 2.4: Maximum clade credibility phylogeny of MP sequences reconstructed under symmetrical phylogeographical model Clades are coloured according to sample location. The MRCA of this segment is inferred to have existed in Asia region, and according to the tMRCA estimates; the virus might have been in circulation for few before its detection as pandemic**.**

2.3.2.2 POPULATION DYNAMICS OF THE 2009 INFLUENZA A/H1N1 PANDEMIC VIRUS IN AFRICA

The population dynamics of Influenza A/H1N1 virus between 2007 and 2014 were investigated using the coalescent Bayesian skyline plot (BSP) model implemented in BEASTv1.8.2. A general trend of rapid virus population growth was seen during early phases of the epidemic in 2009 (figures 5 followed by a period of slow population growth between 2012 and 2014. These trends coincide with the peak of the epidemic period between 2009-2010; followed by persistence of the virus between 2011 and 2014. Thereafter the pandemic was declared by the WHO to be over.

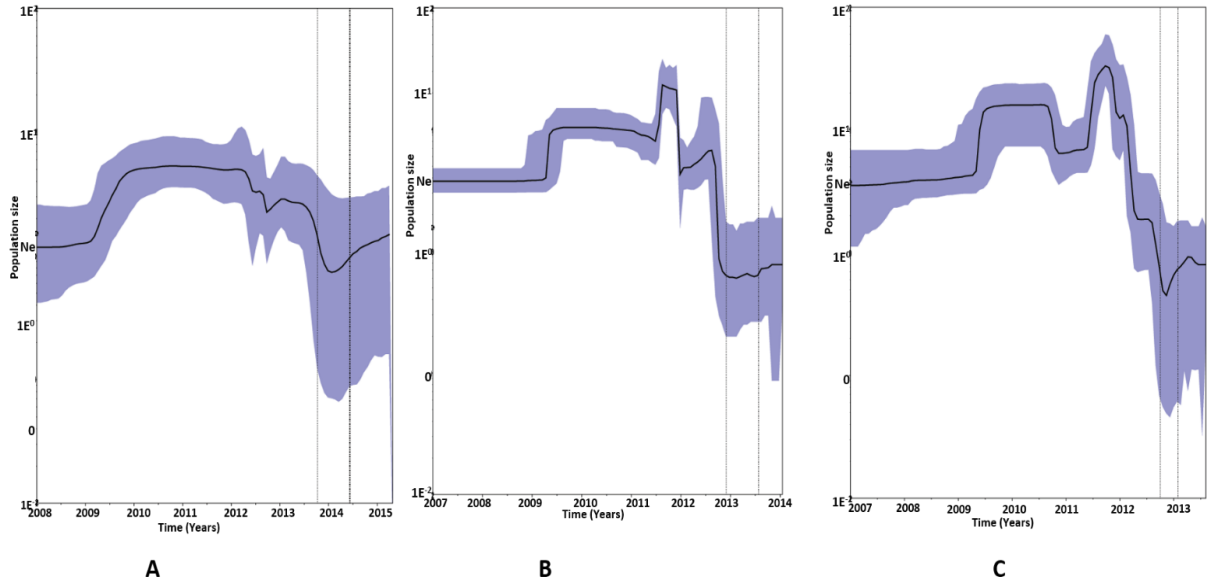


Figure 2.5: Population dynamics of the 2009 Influenza A/H1N1 pandemic virus

Panel A: HA segment sequences. The X-axis shows time in years between 2007 and 2014 (sampling period 2009-2014) while the Y-axis represents the effective population size (Ne). The thick solid line represents the median population over time whereas the upper and lower margins of the blue area represent the 95% highest probability density of the Bayesian skyline estimation of the relative effective population size (Ne**). Panel B:** NA segment sequences. X-axis shows time in years before mid-2014 (sampling period 2009-2014) while Y-axis represents the effective population size (Ne). The thick solid line represents the median population over time whereas the upper and lower higher margins of the solid area represents the 95% HPD of the Bayesian skyline estimation of Ne over the sampling period. **Panel C:** MP segment sequences. X-axis shows time in years before mid-2014 (sampling period 2009-2014) while Y-axis represents the effective population size (Ne). The thick solid line represents the median population over time whereas the upper and lower higher margins of the solid area represents the 95% HPD of the Bayesian skyline estimation of Ne over the sampling period

2.3.3 Spatial origins of 2009 the Influenza A/H1N1 pandemic virus in Africa

When a new or a known but known pathogen emerges, important steps towards its mitigation are: (1) tracing it back to its geographical origin and, (2) retracing the directions and timings of the movement events that have yielded the pathogen's present observed distribution. By

using the BSSVS method implemented in BEASTv1.8.2 under a symmetrical diffusion model, the most probable geographical sources of the H1N1 variants that reached the African continent were inferred.

According to the discrete phylogeographic analysis using HA sequences, the most probable ancestral location for this segment of the virus was in Asia. Using NA sequences, under the same model the ancestral location for the these segment of thevirus under investigation was North America (Figures 2-3). The geographic origin as suggested from the discrete phylogeography analysis using MP sequences was Asia (Figure 4). The fact that different geographic origins were inferred using data from the three different segments is in line with the mixed origin of the genetic segments that make up this virus i.e. it is was demonstrated to be a triple reassortant virus. The symmetrical social network discrete phylogeography analysis which considers movements of viruses between any two pairs of locations showed eleven significantly supported pathways as evidenced by associated Bayes Factors ≥ 3 (Table 2.4). Highly supported intra-Africa movements were observed between eastern Africa and western Africa, central and western Africa, eastern and southern Africa, eastern and northern Africa (Table 2.4). A few location pairs between Africa and non-Africa regions were inferred to have experienced virus exchanged based on significant Bayes factor support e.g. between Eastern Africa and Oceania, Asia and Eastern Africa, Eastern Africa and Europe, North America and Western Africa as well as between Eastern Africa and South America (Table 2.4).

FTable 2.4: Significant movement pathways between pairs of location inferred from discrete phylogeography analysis of MP sequences

Dispersal pathway: Between		**Bayes Factor (BF)**
Eastern Africa	Western Africa	24317.03
Eastern Africa	Oceania	24317.03
Central Africa	Western Africa	24317.03
Asia	Eastern Africa	6076.52
Eastern Africa	Europe	325.02
Eastern Africa	Southern Africa	221.54
Asia	Northern America	166.43
Eastern Africa	Northern Africa	149.32
North America	Western Africa	75.84
Asia	Europe	9.98
Eastern Africa	South America	8.597

2.3.4 SPATIOTEMPORAL DIFFUSION PATTERNS 2009 INFLUENZA A/H1N1 PANDEMIC VIRUS IN AFRICA

The continuous dispersal of the 2009 Influenza A/H1N1 pandemic virus in Africa in unobserved locations was explored using both a homogenous diffusion model and a suite of relaxed diffusion models (Cauchy, gamma and lognormal) which allow rates of diffusion to vary in time and space. The gamma diffusion model was better supported for the HA and NA datasets whereas the homogeneous diffusion model was better supported for the MP dataset. This method of model testing is based on posterior probabilities and best supported models for each dataset are shown with AICM scores (in bold) as provided in Tracer v1.6 (Table 2.5).

Table 2. 5: Continuous phylogeograhy model test using the AICM method for HA, NA, MP datasets

MP	AICM	S.E	cauchy	homogen	lognormal	Gamma
Cauchy	12273.696	+/- 15.768	-	-383.137	1730.491	57886.57
Homogenous	**11890.559**	+/- 2.119	383.137	-	2113.628	58269.707
Lognormal	14004.187	+/- 5.218	-1730.491	-2113.628	-	56156.079
Gamma	70160.266	+/- 13.621	-57886.57	-58269.707	-56156.079	-

NA	AICM	S.E	Cauchy	Gamma	Hom	Lognormal
Cauchy	31904.509	+/- 15.959	-	-2781.009	-983.96	-161.659
Gamma	**29123.5**	+/- 29.055	2781.009	-	1797.049	2619.35
Homogenous	30920.549	+/- 11.99	983.96	-1797.049	-	822.301
Lognormal	31742.85	+/- 7.926	161.659	-2619.35	-822.301	-

HA	AICM	S.E	Cauchy	Gamma	Hom	Lognormal
Cauchy	36390.24	+/- 14.666	-	-1291.603	465.459	2655.624
Gamma	**35098.638**	+/- 10.981	1291.603	-	1757.062	3947.226
Homogenous	36855.699	+/- 7.328	-465.459	-1757.062	-	2190.165
Lognormal	39045.864	+/- 29.85	-2655.624	-3947.226	-2190.165	-

The predicted rate of dispersal was 1351km per year. This points to approximately 3.7km per day With Eastern Africa predicted to have played a crucial role in disseminating the virus to other parts of the continent.

2.3.5 PREDICTORS OF 2009 INFLUENZA A/H1N1 PANDEMIC VIRUS SPREAD IN AFRICA

In this analysis, several factors thought to contribute to the spread of the 2009 Influenza A/H1N1 pandemic virus in Africa were tested using Bayesian Stochastic Search Variable Selection (BSSVS) implemented in BEASTv1.8.2. BSSVS enables the addition of an indicator variable for each pairwise transition rate that specifies whether the rate is on or off, i.e. at its estimated value i.e. 1 if the predictor is included or at 0 when it is not included in the Generalized linear model (GLM). The indicator variable denoted by delta in the GLM formulation gives the inclusion probability of any predictor whereas the conditional effective size i.e. the contribution of each predictor is denoted by beta in the formulation. The statistical support for each predictor being implicated in explaining the observed dispersal pattern was assessed by calculating the Bayes Factors from the posterior distribution density (Table 2.6). Any predictor whose statistical support in terms of Bayes factor values was ≥ 3 was considered significant.

In the initial analysis, it was observed that the predictor sample size and location latitude contributed significantly to the spread of 2009 A/H1N1 virus in Africa. However as these factors are likely to have been strongly influenced by opportunistic sampling their contribution was not further considered. Similarly, a number of countries lie on the same latitude geographicaly and that would also confound the observation that location latitude played a significant role in the dispersal pattern of the H1N1 virus within the African continent.

The number of air passengers, number of flights, and geographical distances between pairs of locations showed significant support,where as great circle distances had decisive support with a BF of 43434 (Table 2.6).

Table 2.6: Statistical support (Bayes Factors) for each the predictors tested in the GLM model

Predictor	Bayes Factor
donor Agglomeration index	3.42
recipient Agglomeration index	0.29
donor GDP	0.11
recipient GDP	0.35
donor trade exports	0.06
recipient trade exports	0.16
donor trade imports	0.16
recipient trade imports	0.03
donor location long	0.22
recipient location long	0.03
donor no. air passenger	8.79
recipient no. air passenger	7.15
donor no. depart flights	5.40
recipient no. depart flights	12.76
donor incidence	0.22
recipient incidence	0.09
donor no of cars	1.28
recipient no of cars	0.28
donor Pop. Size	0.18
recipient Pop. Size	0.09
donor railway coverage	0.08
recipient railway coverage	0.04
donor Vac Coverage	1.56
recipient Vac Coverage	0.03
Great Circle Dist	43434.78
Road Dist	0.66

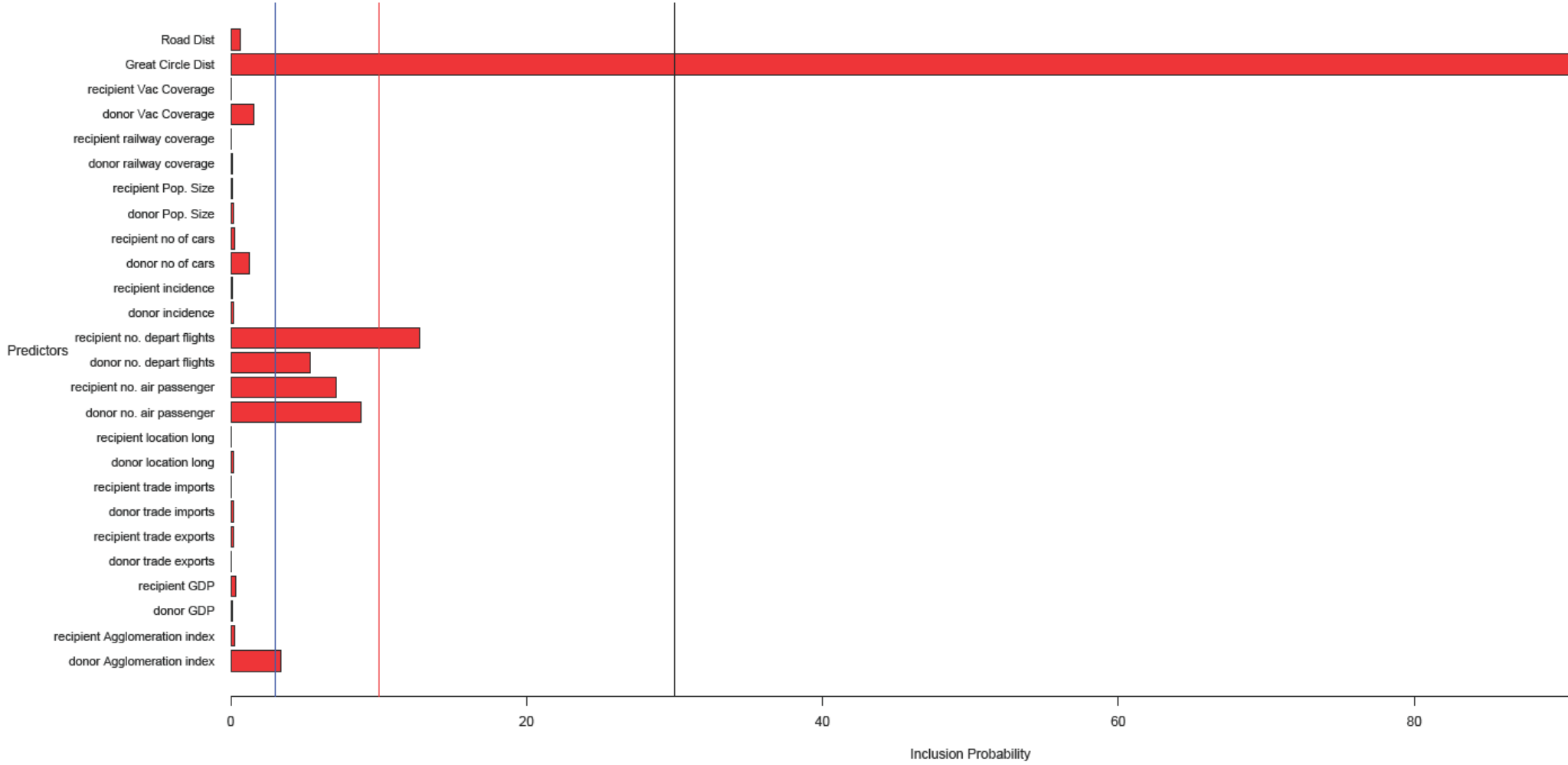


Figure 2.6: Inclusion probability (frequency) of the potential predictors tested for their contribution to the observed pattern of spread of Influenza A/H1N1 pandemic virus in Africa

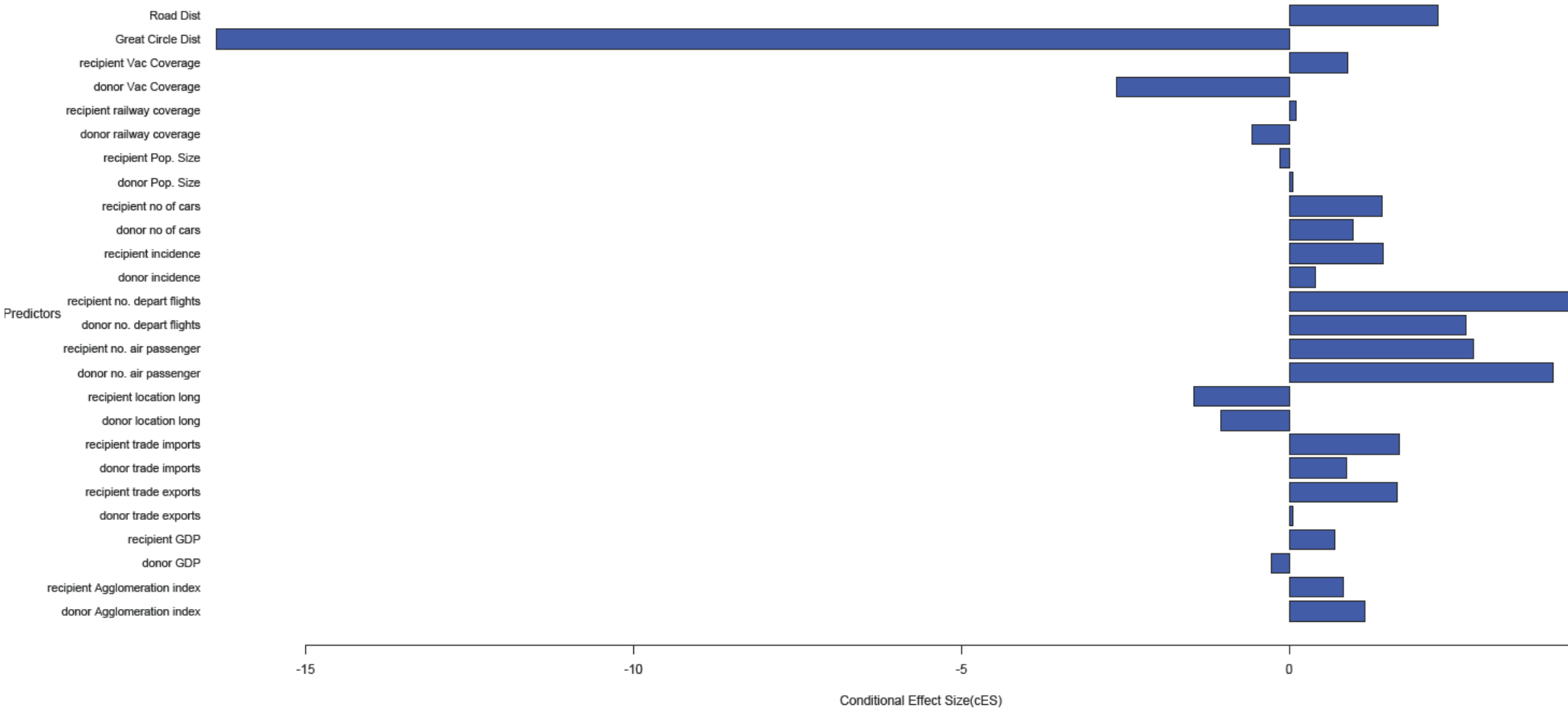


Figure 2.7: Inferred Predictor contribution to the observed spread as estimated from GLM analysis: The x-axis indicates the predictor while the y-axis shows the GLM coefficients which quantifies the conditional effect Size of each predictor included in the analysis. Latitude, Sample size (number of sequences per sampled location) and great circle distances contributed significantly to the spread of 2009 H1N1 virus in Africa

2.4 DISCUSSION

In this study, the epidemic profile of the 2009 Influenza A/H1N1 pandemic in Africa was studied using HA, NA and MP genetic sequence data and several ecological, economic, demographic and genetic factors that could potentially be driving the observed dispersal pattern were tested and their role quantified using the recently developed GLM model implemented in BEAST (Beard *et al.*, 2014). Sequence data from 26 African countries made available in the GISAID open database was combined with other global data from NCBI Influenza Ressearch database (IRD) to make final sequence datasets that were then used to perform analyses reported herein. The choice of the HA and MP and NA segments for analysis was guided by the fact that membrane associated proteins play key role in driving the infections in humans. They form the first portions of the virus that interact with human cells during infection. Although human immune responses are directed at the proteins encoded by these segments, they are able to acquire mutations so that they can evade immune responses.The principal finding from this study suggests that introduction of the H1N1 pandemic virus to Africa was multifaceted. This is evidenced from multiple distinct African lineages in phylogenies reconstructed from the HA, NA and MP segments. The distinct lineages can be traced to different ancestral locations for each gene (Figures 2.2-2.4). The time line of the introduction also varies slightly although in most regions in Africa the virus was likely present as early as June 2009, shortly after being detected in Mexico and the USA in early 2009. The reconstructions also point to a realization that this virus was in circulation as far back as 2007, which is approximately 2 years before its initial detection in North America. This is evidenced by the time to most recent common ancestor (tMRCA) being approximately seven years before the latest sampling date in mid-2014 for some sequences included in these analyses.

The eastern African region is inferred to be the most common location for the within-Africa dissemination of the virus to other African regions. Factors that may support this relate to the fact that countries in the eastern African region form transport gateways to other African countries. Some of the earliest introductions of this virus into Africa are inferred to have occured via this region (Figures 2.2-2.4). Importantly, despite this region contributing the highest number of samples included in the analysis, it was not inferred as the ancestral location.

Central Africa is seen as the region to have witnessed the latest emergence the virus during the epidemic, although this could be attributed to late initiation and scale-up of surveillance efforts in this region. This finding corroborates that of an earlier study that reported a delayed onset of the 2009 Influenza A/H1N1 pandemic in west and central Africa regions (Nzussouo *et al.*, 2012). The discrete phylogeographic analyses that were conducted here revealed multiple independent introductions of Influenza A/H1N1 viruses into Africa (Figures 2-3). Although not explicitely proven here, this is almost certainly a consequence of human mobility due to air travel.

Evidence of multiple entries of A/H1N1 into Africa during this pandemic is somewhat concerning, because these occurred despite concerted internationally coordinated efforts to control the spread of this virus. A similar study using HA sequences showed evidence of inter-region virus dispersal by reconstruction of past spatial transmission patterns for the 2009-2014 period with annual global epidemic intervals spanning October to September each year, for nine geographic regions (Africa, Australia, East Asia, Europe, the Middle East, North America, South America, South Asia and Southeast Asia). The HA study revealed several decisive and strongly supported introductions into Africa within the study period and included Europe to Africa (2009-2010), North America to Africa (2009-2010), East Asia to Africa (2010-2011), Europe to Africa (2010-2011), South Asia to Africa (2011-2012), Europe to Africa (2012-2013) and East Asia to Africa (2013-2014) (Su *et al.*, 2015a). The current study supports aforementioned seminal study and additionally captures previously unreported introductions from Oceania to Africa and intra-Africa regional exchanges of this virus, thus providing a more regional scale of spatial dynamics of pandemic H1N1 circulation within the African continent during similar period.The three estimates of effective H1N1 population sizes during and after the pandemic are broadly consistent with one another and indicate a trend typical of an outbreak situation (Figures 5). During the onset of the epidemic, an exponential rise in the effective population occurred in a short time period between mid to late 2009. This observation corroborates well with the observed spread pattern in the population given it was a novel strain of the Influenza A/H1N1 virus. Subsequent periods show a plateau phase of the epidemic suggesting a concerted effort of countries responding to it to prevent further spread. This could also have been a result of herd immunity emerging from the population. In the time following the epidemic (i.e. 2013-2014), there is decline in the effective population size as the virus became just another seasonal Influenza A/H1N1 virus.

During the initial phase of an epidemic, there is an expectation that most individuals in the population will be susceptible to a newly emergant virus strain. Therefore, any counter measures taken during this initial phase should be maximally impactful with respect to containing an outbreak. If they are made early enough, it is hoped that interventions such as the deployment of vaccines, the implementation of travel restrictions, and quarantine measures could effectively prevent a regional outbreak from becoming a global epidemic.

The 2009 Influenza A H1N1 pandemic was genetically characterized and found to be composed of geneome segments derived from human, bird and swine viruses. The different spatial origins of the HA, NA and MP segments inferred in this study is in line with the finding that this virus had mixed geographical ancestry (Garten *et al.*, 2009; Trifonov *et al.*, 2009). The HA sequences analysed here were lineages circulating in swine populations. A study by Xu et 2012 investigated host specific differences in the HA-NA functional balance in human and swine derived pandemic H1N1 viruses and found a striking difference between these. Specifically they demonstrated that the balance is conserved in human viruses whereas it is lacking in swine-derived viruses. This suggests that swine viruses might have acquired this characteristic prior to transmission to humans in the pre-pandemic periods (Xu et al 2012). This is supported by the current analysis where the HA phylogeny indicated that the ancesstors of the pandemic strain were already circulating in swine two years prior to detection in humans in early 2009.

Conversely the MP gene is inferred to have geographically originated from Asia suggesting a close relationship with avian H1N1 lineages and support the inferences made by (Garten *et al.*, 2009). The discrete phylogeographic analysis of NA revealed that its most recent common ancestor existed within Eurasian lineages and contrasts the findings in the current analysis that show that NA sequences have their geographical origins in North American territories.

Following the initial outbreak in Mexico it took less than 3 months for A/H1N1 to reach all continents of the world (Widdowson, Iuliano and Dawood, 2014). Continuous phylogeographic analysis revealed a fairly fast rate of spatial spread of over 100 kilometers per day. The direction of the dispersal appears to have been from Asia, North America or Europe into Africa. Since the model used considered scenarios of reversible movement, the possibility of out of Africa movements of viral variants is also plausible.

Several factors could fuel a global pandemic of this magnitude. In Africa, as with the rest of the world, a mixture of genetic, ecological and economic factor likely contribute to the rapid spread of infections. To explore and perhaps unravel critical factors impacting on the transmission of A/H1N1 in Africa, the newly developed Generalized Linear model (GLM) implemented in Bayesian phylogenetic tool namely BEAST was employed as previously applied in a number of other recent studies (Lemey *et al.*, 2014; Magee *et al.*, 2014; Nunes *et al.*, 2014; Gräf, Vrancken, Maletich Junqueira, *et al.*, 2015) . Degrees of support for the various ecological, economic, and genetic predictors of virus sread were assessed from Bayes factors and the posterior inclusion probabilities obtained following the dicrete/continuus phylogeographic analyses performed here. Collectively, 16 predictors were analyzed. Any predictor having an inclusion posterior probability of 30% and higher was deemed to explain in part the diffusion of the virus within the African region.

The initial hypothesis that road connectivity could have played key role in the dissemination of A/H1N1 in African was rejected in our GLM analysis. Instead, simple geographical distance was inferred to be the spatial factor that made the most significant contribution to the observed dynamics of A/H1N1 spread within Africa.

The finding that geographical distances and predictors associated with air transport strongly impact the spatial diffusion of virus is consistent with previous studies which have demonstrated that environmental factors such as temperature, humidity and longitude as well as human mobility have an impact on the spatial epidemiology of Influenza viruses (He *et al.*, 2013; Lemey *et al.*, 2014; Magee *et al.*, 2014). Although the rapid diffusion of Influenza viruses globally has been attributed to air travel (Lemey *et al.*, 2014), and this mode of transport is almost certainly the mechanism by which the virus was introduced to Africa from the outside world, it is likely that movements of the virus within Africa are not primarily driven by this mode of transport. This could be explained by the infrequent use of air travel by most people when travelling within Africa. For example, a predominance of road mediated transmission has been identified in analyses of rabies virus in North Africa (Talbi *et al.*, 2010) and avian flu in Nigeria (Rivas *et al.*, 2010).

Influenza surveillance and research activities have been scaled-up in recent years mostly driven by the emergence of influenza strains originating in avians and swine. Several African countries are now including influenza in their routine infectious disease surveillance programs. However, the full potential of these programs have yet to be realized and, as a consequence,

each year only a small fraction of the influenza genome sequence data that is generated originates from Africa. A major limitation of the study described here is both that the amount of sequence data from Africa is lower than that in the rest of the world, and that the African sequences that are available have been opportunistically sampled. As a consequence of tehse two factors it is possible that inferences about the H1N1 epidemic that were made with these sequences do not perfectly reflect what actually occurred accross the entire continent during the epidemic.

2.5 CONCLUSION

The ability to predict what will happen in future epidemics based on what has happened in past epidemics is for a crucial component of influenza prevention and control, and should be especially important in an African context where resource constraints demand that infection control be carried out with maximum efficiency. In our globalised, interconnected world local outbreaks of infectious disease can quickly become global challenges. The particular challenges encountered in different regons of the world, even when considering the same pandemic, will require individualized country or region specific solutions. It is hoped that by highligting the imprtnance of local spread and nort-south dispersal this study will in some small way improve the ability of African countries to effectively respond to such epidemic. With current model modification and improvements to include negative controls that may further refine the predictions of virus spread.

CHAPTER 3

REASSORTMENT AND SPATIAL DYNAMICS OF INFLUENZA VIRUSES

3.1. INTRODUCTION

Influenza viruses form the three biggest taxonomic grouping within the family of *Orthomyxoviridae* (Lamb, Krug and Knipe, 2001). As with other orthomyxoviruses they are negative sense segmented RNA genomes. They are broadly classified into three types: Influenza A, B and C (Forrest and Webster, 2010). Influenza A and B are characterized by eight individual gene segments that make up their genomes whereas the Influenza C genome spans seven gene segments. The natural reservoirs of Influenza viruses are wild aquatic birds such as waterfowls (Webster *et al.*, 1992). Influenza A viruses are the most diverse of the influenza viruses and are known to have a broad host range, whereas Influenza B and C viruses have been mostly isolated from humans and a few other non-human hosts (Kobasa and Kawaoka, 2005; Merler *et al.*, 2011; Bahl *et al.*, 2013).

Segmented viruses evolve through two major mechanisms: mutation and re-assortment (Bedford *et al.*, 2014; Neher, Russell and Shraiman, 2014; Steel and Lowen, 2014; Dudas *et al.*, 2015). Whereas the former is attributed to the error prone virus-encoded RNA dependent RNA polymerase enzyme that is required for genome replication, the latter occurs as a result of super-infection of individual cells with two or more distinct strains of a particular type of virus that may then exchange segments during the course of replication and encaspsidation (Zhou *et al.*, 1999; Lycett *et al.*, 2012; Tao, Steel and Lowen, 2014; Bedford *et al.*, 2015).

Re-assortment is considered a special form of genetic recombination (McDonald *et al.*, 2016). A number of experiments geared towards understanding the underlying causes and consequences of reassortment in segmented viruses such as the Influenza viruses have been conducted to date and these have revealed that reassortment is an ongoing process and usually occurs in mixed infections with viruses with compatible genomic structure ie any mismatch in the segments reduces the probability of reassortment (Desselberger *et al.*, 1978; Gao and Palese, 2009; Boni *et al.*, 2010; Marshall *et al.*, 2013; Tao, Steel and Lowen, 2014). Additionally, with the rapidly growing availability of large volumes of full-genome sequence data and the development of computational tools to study recombination and reassortment in

such data, *in silico* analyses can now reveal a great deal about patterns of reassortment in natural virus populations (Yurovsky and Moret, 2010; Nagarajan and Kingsford, 2011; T. T. Y. Lam *et al.*, 2013; Lu, Lycett and Brown, 2014; Martin *et al.*, 2015). Although reassortment is thought to occur in almost all segmented RNA viruses, it has so-far mostly been studied in Influenza A viruses and there remains little data on this process in Influenza B and C viruses (Boni *et al.*, 2010; Lam *et al.*, 2011; Nagarajan and Kingsford, 2011; Neverov *et al.*, 2014; Westgeest *et al.*, 2014; Dudas *et al.*, 2015; Freire, Iamarino, Soumaré, Faye, Sall, Guan, *et al.*, 2015). Computational analyses therefore also provide the valuable opportunity to compare and contrast reassortment patterns in these other Influenza viruses to those that are found in Influenza A virus, as this has only been previously studied *in vitro* (Baker *et al.*, 2014).

The earliest confirmed historical case of Influenza A virus reassortment, and probably the best example of the potentially devastating epidemiological effects that reassortment can have, relate to the 'Spanish flu' pandemic that occurred between 1918 and 1919. This pandemic has been attributed to the subtype, A/H1N1 (Taubenberger and Morens, 2006; Worobey, G.-Z. Han and Rambaut, 2014). Several studies have indicated that this virus was a reassortant of human and avian Influenza viruses. The epidemic caused by this reassortant was, in terms of morbidity and mortality, the biggest in human history with approximately 500 million cases and 20-50 million deaths occurring worldwide (J. Taubenberger et al., 2005; J. K. Taubenberger & Morens, 2006 ;Worobey, Han, & Rambaut, 2014).

The epidemiological importance of reassortment is underlined by the fact that it is the basis upon which Influenza A viruses have been classified into major lineages such as the classic swine, triple reassortant and Eurasian avian lineages.

Besides the already mentioned A/H1N1 subtype (the one that caused the Spanish flu), the other Influenza A subtypes are also characterized by high frequencies of point mutations and reassortment. A study on the origin of the human Influenza A/H2N2 and A/H3N2 strains suggested that reassortment occurrs at high frequencies in these subtypes and has gaven rise to the currently circulating strains (Scholtissek, 1994). Additionally, Holmes et al., (2005) demonstrated high frequencies of reassortment in Influenza A/H3N2 viruses circulating in the USA (Holmes *et al.*, 2005). This was achieved through large-scale phylogenetic analysis of whole genome sequences. Focusing on avian Influenza A strains, Lu. et al (2014) reported significant amounts of reassortment in the internal genes of avian Influenza A strains through phylodynamics analysis of large nucleotide sequence datasets derived from avian hosts (Lu,

Lycett and Brown, 2014). Lam et. al (2013) also reported large numbers of mosaic genomes through analyses of Influenza virus full genome sequences (T. T. Y. Lam *et al.*, 2013). Similarly, a study by Westgeest et al 2014 revealed high degrees of reassortment in Influenza A/H3N2 viruses isolated in the United Kingdom (Westgeest *et al.*, 2014).

In contrast to Influenza A viruses, Influenza B viruses exhibit more limited reassortment, although various studies have suggested that reassortment could occur at high frequencies between the two major lineages (Victoria and Yamagata) that co-circulate each year (Goodacre, 2013; Dudas *et al.*, 2015; Oong *et al.*, 2015; Suptawiwat *et al.*, 2017). Dudas et al (2015) through the analysis of whole genome alignments of globally circulating Influenza B viruses demonstrated that segments PB1, PB2 and HA undergo limited reassortment compared to other genome segments. In the same study it was found that the PB1, PB2 and HA segments maintained their parental ancestry of either Victoria and Yamagata lineages compared to other genome segments which reasserted frequently and lost their distinct ancestral diversity (Dudas *et al.*, 2015).

Reassortment is also likely an important mechanism in the epidemiology and evolution of Influenza C viruss (Matsuzaki *et al.*, 2003). It has been postulated that it plays a key role in the perpetuation of this virus in humans. Since the first Influenza C virus isolates were identified in 1947, this virus has been circulating in humans and is especially prevalent in Japan where it causes seasonal Influenza epidemics (Pachler and Vlasak, no date; Matsuzaki *et al.*, 2003; Ludwig, 2014). The six main Influenza C virus genetic lineages have been classified based on sequence analyses of the HEF segment and include: Yamagata/26/81, Aichi/1/81, Mississippi/80 Taylor/1233/47, Sao Paulo/378/82 and Kanagawa/1/76 (Speranskaya *et al.*, 2012). In a 9-year survey carried out in japan, Matsuzaki et al (2016) detected that reassortment amongst Influenza C viruses coupled with immune selection resulted in a variety of different viral strains each comprising different collections of genomic segments that had been derived from the reference lineages (Matsuzaki *et al.*, 2016).

Most reassortment studies in Influenza viruses have been conducted empirically *in vitro* through co-infection of cell lines, or *in vivo* through co infection of model organisms with pairs of viral strains followed by assessments of whether reassortment occurs. Although the earliest these studies simply demonstrated that reassortment in Influenza viruses was possible (Desselberger *et al.*, 1978), more recent studies have revealed that reassortment is a nonrandom process that is guided by segment-specific packaging signals. One study observed that avian

origin segments HA, M and PA were linked during exchange as demonstrated by the genomic anatomy of the resulting virions (Essere et al., 2013). Another experimental reassortment study revealed that intertype Influenza virus reassortment may occur in the presence of compatible viral packaging signals when they were able to generate viable virions from Influenza A and B segments through the addition of Influenza A virus packaging signals to full length Influenza B virus glycoproteins (Baker *et al.*, 2014). However, a similar study indicated that heterologous packaging signals in segment 6 and 8 failed to limit Influenza A virus reassortment whereas segment 4 (HA) showed limited reassortment in the presence of heterologous packaging signal constellations (White, Steel and Lowen, 2017). Collectively, these studies suggest that segments that are genetically compatibile with respect to their segment specific packaging signals, will be interchangeable during reassortment events .

The number of segments in a segmented viral genome has a large impact on how effectlively reassortment to generate novel genetic variants (McDonald *et al.*, 2016; White, Steel and Lowen, 2017). For example Influenza A and B viruses each have eight segments and it is therefore possible for reassortment to yield 256 (or 2^8) genetically unique reassortants from any two parental genomes, whereas reassortment between two Influenza C genomes – each of which contain only seven components – only has the potential to yield 128 (2^7) genetically unique reassortant progeny genomes.

In this chapter, I set out to test two hypotheses: (1) that there is no difference in the frequency of reassortments among the segments that constitute influenza virus genomes; and (2) that there is epochal temporal reassortant among influenza viruses and that all geographical regions are equally likely sources of epidemiologically important influenza virus reassortant lineages. To achieve this, I describe an exploratory analysis of reassortment in Influenza A, B and C full genome sequences that have been sampled over eight decades between 1927 and 2014. I use the time when, and geographical locations where, these sequences have been sampled to phylogeographically reconstruct the temporal and geographic origins of reassortant lineages. These analyses yield a more detailed understanding of how reassortment has impacted the emergence and epidemiology of Influenza lineages over the past century.

3.2 MATERIALS AND METHODS

3.2.1 SEQUENCE PREPARATION

I downloaded all full or near-full genome length sequence data for Influenza A (4200 sequences), B (1800 sequences) and C (150 sequences) viruses from the NCBI Influenza Virus

Resource database through its FTP server (available at ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/). I additionally downloaded 200 Influenza A, 120 Influenza B and 35 Influenza C sequences from the GISAID public sequence repository (https://www.gisaid.org/) (Elbe and Buckland-Merrett, 2017). All of these sequences were sampled between 1927 and 2014.

The criteria for inclusion of sequences in my final datasets were that: (1) they should be either full or near full length (i.e. representing 80-90% coverage of the full length size of the respective segment); (2) the date of collection was clearly provided; and (3) the isolate or strain had all the genomic segments sequenced (i.e. eight segments for Influenza A and B or seven segments for Influenza C). Following these criteria, I additionally removed incomplete genomes and duplicated segment sequences. Sequences for each segment were aligned separately using Muscle (Robert C Edgar, 2004). Whole genome sequence alignments were assembled by concatenating these individual segment alignments. These full genome alignments were then used for reassortment and phylogeographic analyses.

3.2.2 Reassortment analysis

I performed an exploratory analysis to uncover the presence and the extent of reassortment in three datasets representing concatenated full genome sequences of Influenza A, B and C viruses. This was achieved using RDP4.62 with default settings (Martin *et al.*, 2015). RDP4.62 implements seven recombination detection methods: RDP, MAXCHI, BOOTSCAN, CHIMAERA, 3SEQ, SISCAN and GENECONV). Potential reassortment events that were detected by three or more of the different recombination detection methods implemented in RDP4.62 with a Bonferroni corrected p-value cut-off of 0.05, and with phylogenetic support for the occurrence of reassortment and/or recombination, and where detected recombination breakpoints fell at the interfaces between segments, were taken as representing evidence of genuine reassortment events.

3.2.3 Evolutionary and temporal Analysis to determine where and when reassortment likely occured

Given the computational intensity of the Bayesian phylogenetic analyses, a computationally feasible dataset had to be selected from the larger alignments used for the reassortment analyses. Representative sequences from each of the three full genome alignments of Influenza

A, B and C were selected using Usearch (Edgar, 2010) with a 90% identity cut-off. This yielded representative alignments respectively containing 315, 200 and 97 whole Influenza A, B and C genome sequences. Using these representative alignments, I performed a reassortment analysis following the same strategy and settings as previously described in the reassortment analysis section (see section on reassortment analysis). The dates when and locations where isolates were sampled are useful for reconstructing dated phylogenies with the aim of inferring the spatiotemporal dynamics of the putative reassortant lineages. I first tested the 'clock-likeness' or temporal signal within these "time-stamped" datasets using TempEst (formerly Path-o-gen) (Rambaut *et al.*, 2016) in order to determine the appropriateness of the molecular clock assumption when reconstructing time-scaled phylogenies in BEAST v1.8.2 (Drummond and Rambaut, 2007). I tested the best fit clock model for the data and found that the relaxed clock model best supported the three datasets (Drummond *et al.*, 2006). Since I was interested in identifying the ancestral lineages that gave rise to the current reassortant lineages the coalescent Bayesian skyline model was used so as to enable inferences of both the geographical and temporal origins of these ancestral sequences while at the same time accounting for the complex seasonal population dynamics of Influenza viruses (Drummond, Rambaut and Xie, 2011). To estimate the probable times of reassortment for each individual reassortment event, I first identified the node in the MCC tree that represented the most recent common ancestor (MRCA) of the putative reassortant and corresponding predicted major and /or minor parental sequences. The inferred date and the upper bound of the 95% Higher Posterior Density (HPD) for this node were respectively taken as being the upper bounds and upper 95% HPD bound of the date when reassortment event occurred. The lower bound and lower 95% HPD bound of the date when the reassortment event occurred was estimated from the dates and lower bound of the 95% HPD for the node immediately ancestral to the MRCA node in the MCC tree (Tee *et al.*, 2009; Tongo *et al.*, 2018).

3.2.4 DISCRETE PHYLOGEOGRAPHIC ANALYSES

Each sequence in the three datasets was geocoded according to geographical region of isolation at both continental and sub-continental levels. Each sequence in the Influenza A alignment was assigned to one of eleven geographical areas namely: Africa (n=50), Asia (n=43), Europe (n=37), North America i.e. Canada, Mexico, USA (n=79) , Oceania (n=56) and South America (n=50). Each of the influenza B sequences was assigned to one of 6 geographical locations: Africa (n=4), Asia (n=55), Europe (n=11), North America i.e. Canada, Mexico, USA (n=65) , Oceania (n=60) and South America (n=5). Each of the Influenza C sequences was assigned

to one of eight geographical locations: Australia, Brazil, India, Japan, Philippines, Singapore, South Africa and USA.

The phylogeographic analyses were carried out under a Bayesian Stochastic Search variable selection (BSSVS) method implemented in the program, BEAST. The BSSVS method allows for the integration of model uncertainty into the predictions made by the analysis through averaging over models weighted by their relative plausibility; an analytical approach which usually improves predictions of ancestral sequence geographical locations (P Lemey *et al.*, 2009). Additionally, an asymmetrical location-state substitution model was used which enabled inference of the polarities of movements between locations. I inferred possible locations where reassortment events likely occurred by taking the mean root state probabilities of the MRCA node of related reassortant and the node immediately ancestral to this MRCA node that bounded the branch along which the reassortment event was predicted to have occurred.

3.3 RESULTS

3.3.1 DATASETS

To study the reassortment pattern in the three types of Influenza viruses, full genome sequences were prepared by concatenating respective segments of their multipartite genomes. Towards this end, I generated final alignments comprising of 3000, 1302 and 97 full genome sequences of Influenza A, B and C, respectively. These alignments were subsequently used for reassortment analyses.

Although the sequences within these large alignments were sampled throughout the world, some of the "richer" regions (such as North America and Europe) were represented in the databases by higher numbers of sequences; the "poorer" regions (such as Africa and South America) were represented by fewer sequences. This bias needed to be accounted for in many of the subsequent model-based analyses by subsampling of sequences from the full dataset so as to achieve datasets containing a more even spread of sequences.

Using the Usearch clustering tool to sub-sample spatially unbiased datasets from the full dataset, I constructed Influenza A, B and C datasets respectively containing 263, 205 and 97 sequences. Further balancing of the spatial distribution of sequences was achieved by defining geographical locations such that all locations had approximately the same numbers of associated sequences. This anti-biasing was necessary as most models used in these

phylogeographic analyses are sensitive to the sampling scheme: i.e. the sampling scheme can influence inferences made from such analyses.

For inference of the ancestral locations of reassortant lineages and their movement dynamics over a period of eight decades between 1927 and 2014, I further carried out analyses on individual segments in the three datasets. This was informed by the understanding that, due to reassortment, there were likely to be differential frequencies of exchange between the different segments of these viruses.

Table 3.1: Genome sequence set that were concatetenated to make whole genome alignments of Influenza A, B and C viruses.

Segment	Coding region	Length (bp)	Position in the alignment	No of sequences
Influenza A virus				
1	PB2	2333	1-2333	3838
2	PB1	2325	2334-4657	3838
3	PA	2214	4658-6870	3838
4	HA	1777	6871-8646	3838
5	NP	1552	8647-10197	3838
6	NA	1482	10198-11678	3838
7	M1, M2	1003	11679-12680	3838
8	NS1, NS2	872	12681-13551	3838
Influenza B virus				
1	PB2	2348	1-2348	1341
2	PB1	2359	2349-4707	1341
3	PA	2251	4708-6958	1341
4	HA	1840	6959-8798	1341
5	NP	1803	8799-10601	1341
6	NA	1526	10602-12127	1341
7	M1, M2	1153	12128-13280	1341
8	NS1, NS2	1041	13281-14321	1341

Influenza C virus				
1	PB2	2314	1-2314	97
2	PB1	2324	2315-4628	97
3	P3	2135	4629-6763	97
4	HE	2010	6764-8773	97
5	NP	1737	8774-10510	97
6	M1, M2	1155	10511-11665	97
7	NS1, NS2	909	11666-12574	97

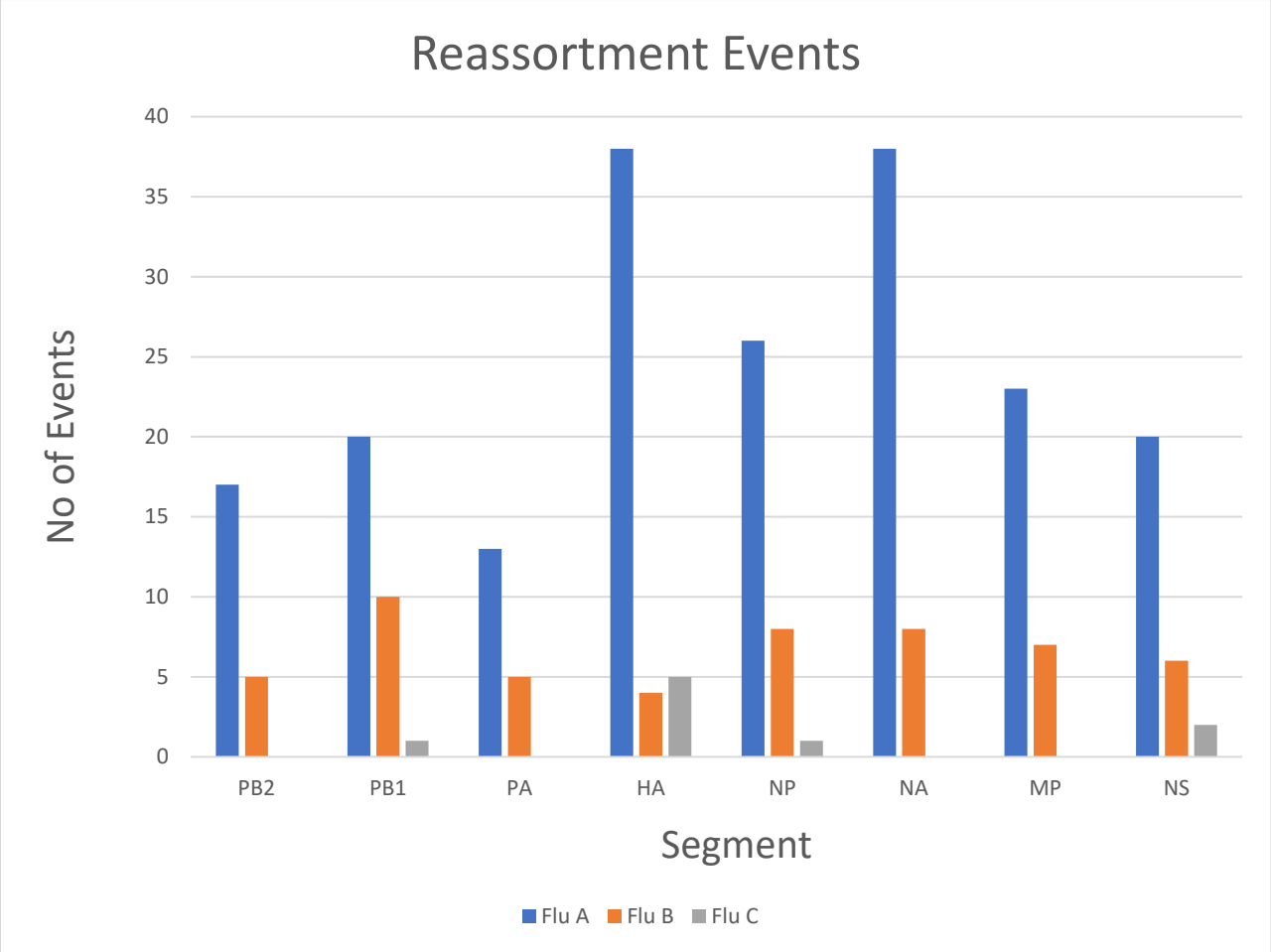


Figure 3.1 Distribution of the number of predicted reassortment events in Influenza A B and C virus datasets per segment along the genome. Influenza A has most events predicted involving HA and NA segments. Influenza B has more events involving PB1, NP and NA segments while Influenza C has majority of predicted events involving HA(HE) segment.

3.3.2 Frequency and patterns of reassortment in Influenza A viruses

Whole genome screening of a total of 3838 Influenza type A virus sequences representing nearly all subtypes using the various reassortment event detection methods implemented in RDPv4.63 identified 175 putative re-assortment events occurring both within, and among, various subtypes of the Influenza A viruses. The reassortment events were considered significant if they had an associated Bonferroni-corrected *p-value* of ≤ 0.05. and were supported by both three or more different reassortment detection methods and phylogenetic analyses. The reliability of these methods to correctly identify putative reassortant sequences was affirmed when mixed sequences included in the alignment were identified as reassortants.

The ancestry of the predicted mosaic sequences showed that the parental genetic distance of of HA and NA gene sequences was relatively higher than those of the internal genes (PB2, PB1, PA, NP, M1/M2 and NS1/NS2) (figure 3.1); i.e. relative to other genome segments, the HA and NA segments tended to be inherited during reassortment events between parental genomes that were more genetically different from one another. It was further noted from the analysis that regions with high number of reassortment events had parental sequence genetic distances of 0.5 whereas the less frequently reassorting genomic regions had parental sequence genetic distance ranging from 0.12 to 0.37 (figure 3.1). The higher frequency of HA and NA transfers during reassortment and the greater parental genetic distances during these transfers implies that, relative to other Influenza A virus segments, the HA and NA segments have a more mixed ancestry.

A heterogeneous pattern of intra-segment recombination events was also evident within some individual Influenza A virus segments. Most of these recombination events are predicted within the surface genes: encoded by the HA and NA segments (Figure 3.2). Since recombination is believed to be very rare in Influenza A viruses (Boni *et al.*, 2010; Forrest and Webster, 2010; Chen *et al.*, 2016), these supposed recombination events are likely either misidentified reassortment events or are laboratory constructed segment chimaeras (T. T. Y. Lam *et al.*, 2013).It is noteworthy that in many cases, reassortment events were detected between parental sequences that might have themselves been reassortants. This pattern highlights the fact that reassortment, and hence mixed infections between genetically distinct Influenza A viruses, is likely very common in this group of viruses.

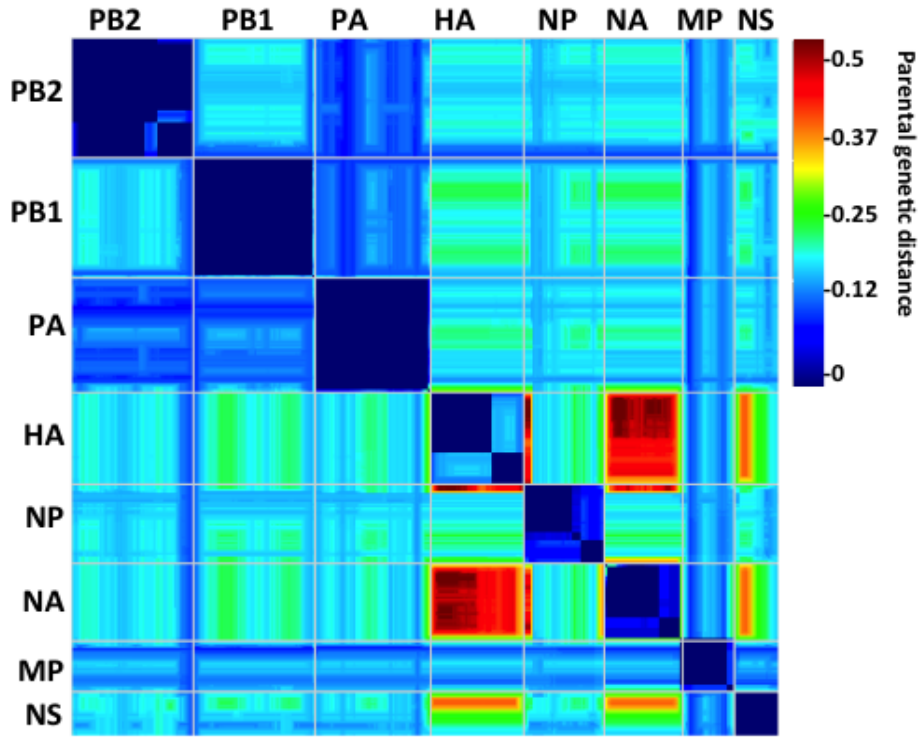


Figure 3.2 **Modularity matrix indicating local Influenza A virus parental sequence genetic distances across the entire genome:** The maximum genetic distances of parental viruses that yielded detectable reassortants/recombinants are depicted here. For any segment the graph should be read either from top to bottom or from left to right. Whereas bluer colours indicate that reassortment events involved transfers of segments into genomes that had low genetic distances (calculated as unweighted Hamming distances) to the genomes from which the segments originated (i.e. the donor and recipient genomes were very similar in blue-coloured genome regions), more red colours indicate that reassortment events involved transfers of segments into genomes that had higher genetic distances to the genomes from which they originated (i.e. the donor and recipient genomes were genetically dissimilar in these genome regions). In reassortment events involving the NA segment, for example, parental virus genome sequences tended to have genetic distances of up to 0.12 for all segments other than the HA and NS segments which, in the case of the HA segment in at least one pair of parental genomes that exchanged an NA segment, had a distance across its entire length of >0.4 between the two parents.

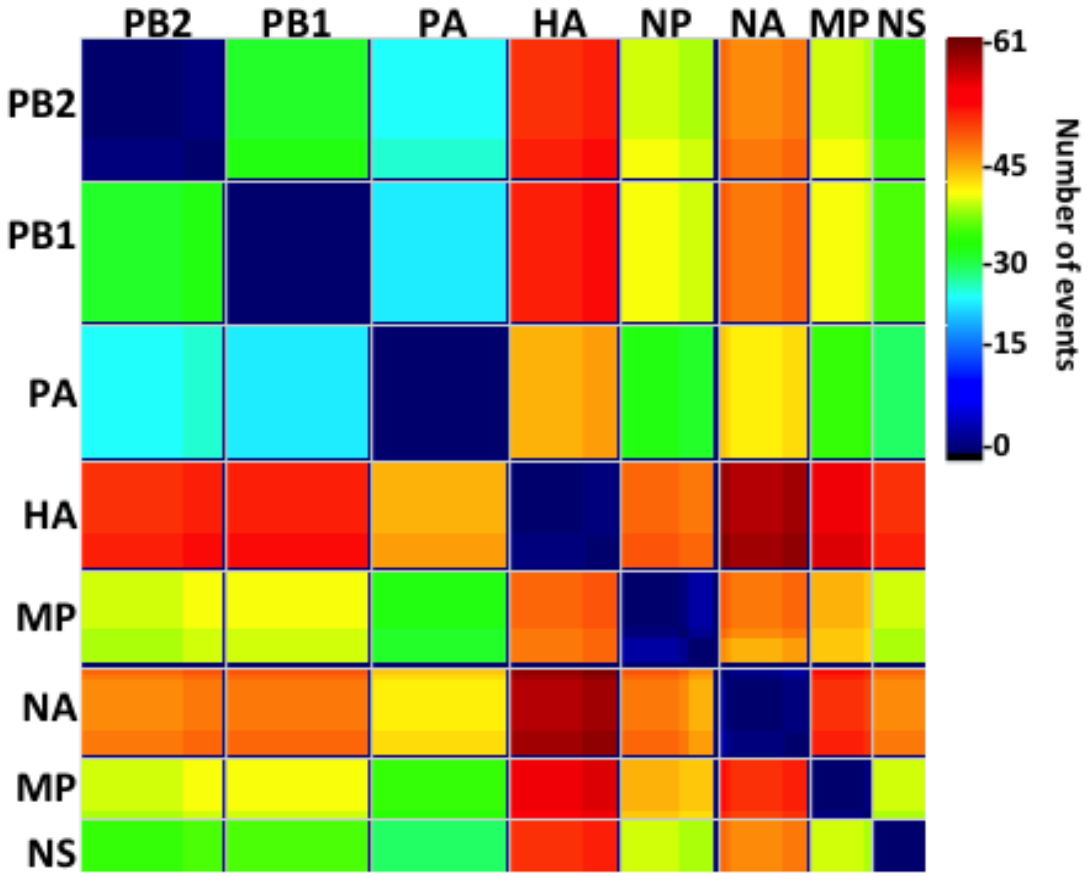


Figure 3.3 **Predicted regional counts of detected reassortment and recombination events** distributed through the genomic segments of the Influenza A viruses. For each segment this plot should be read either from top to bottom or from left to right. With the NA segment row, for example, the yellow/orange/red colors for all segments indicate that whereas reassortment events involving the transfer of NA tend to frequently separate NA from the remainder of segments in the genome (i.e. NA tends to be transferred alone), the frequency with which these exchanges involve the separation of cognate NA and HA segments from one another is particularly high (as is indicated by the red color associated with HA-NA cells in this matrix). Conversely the blue-green colors for PB1-PB2, PBA-PB2 and PB1-PBA cells in the matrix indicates that all of these pairs tended to be inherited together from the same parental genome during reassortment.

3.3.3 FREQUENCY AND PATTERNS OF REASSORTMENT IN INFLUENZA B VIRUSES

Screening of the Influenza B virus sequence alignment revealed a total of 55 putative statistically supported reassortment events. Of these only 40 were detected by three or more of the methods in RDP4.63 together with phylogenetic support (Table 3.3). Most of these events (10/40) involved transfers of the HA and NA segments. However a substantial number of events (8/40, 8/40, 7/40 and 6/40) also involved the transfer of, PB1, NP, MP and NS segments respectively between genetically divergent genomes(Figure 3.3, Table 3.3).

Regional counts of reassortment events ranged from near zero to instances with 17 events having been identified (Figure 3.4). Most reassortment events were detected within segment 6 that codes for neuraminidase (NA) protein. It was also evident that internal gene regions including NP, MP, PB1 and PA had relatively higher reassortment levels given the number of event counts within these regions which were comparatively close to those identified within HA and NA segments (Figure 3.4).

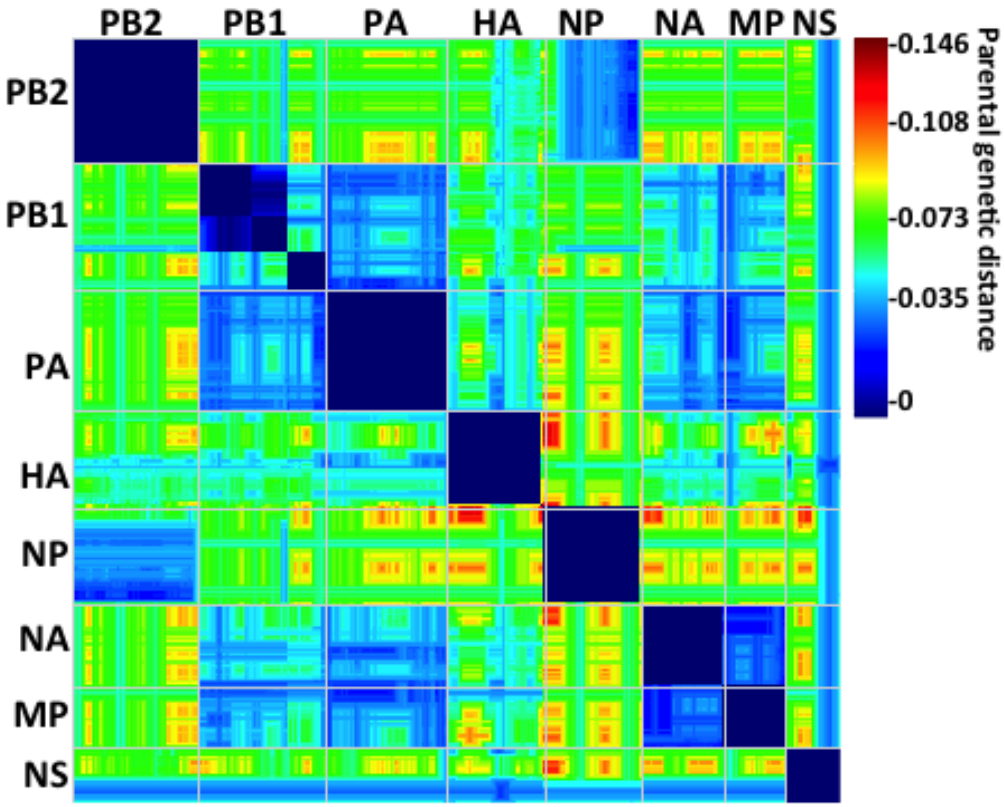


Figure 3.4 **Modularity matrix indicating local Influenza B virus parental sequence genetic distances across the entire genome:** The maximum genetic distances of parental viruses that yielded detectable reassortants/recombinants are depicted here. For any segment the graph should be read either from top to bottom or from left to right. Whereas bluer colours indicate that reassortment events involved transfers of segments into genomes that had low genetic distances (calculated as unweighted Hamming distances) to the genomes from which the segments originated (i.e. the donor and recipient genomes were very similar in blue-coloured genome regions), redder colours indicate that reassortment events involved transfers of segments into genomes that had higher genetic distances to the genomes from which they originated (i.e. the donor and recipient genomes were genetically dissimilar in these genome regions). In reassortment events involving the NA segment, for example, parental virus genome sequences tended to have genetic distances of up to 0.12 for all segments other than the HA and NS segments which, in the case of the HA segment, , had a distance across its entire length of >0.1 between the two parents in at least one pair of parental genomes that exchanged an NA segment.

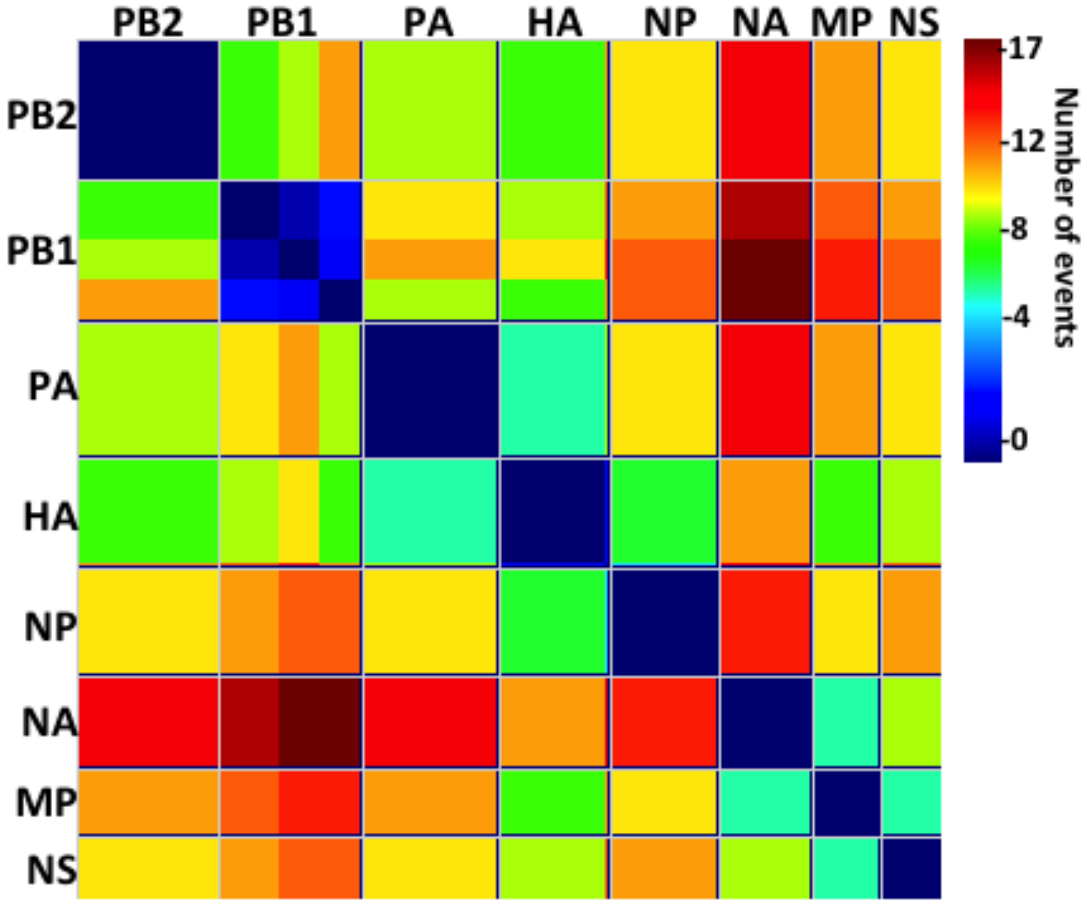


Figure 3.5 **Predicted regional counts of detected reassortment and recombination events** distributed through the genomic segments of the Influenza B virus. For each segment this plot should be read either from top to bottom or from left to right. With the NA segment row, for example, the yellow/orange/red colors for all segments indicate that whereas reassortment events involving the transfer of NA tend to frequently separate NA from the remainder of segments in the genome (i.e. NA tends to be transferred alone), the frequency with which these exchanges involve the separation of cognate NA and PB1 segments from one another is particularly high (as is indicated by the red colour associated with NA-PB1 cells in this matrix), similarly the same is observed between NA-PB2, NA-PA. Conversely the blue-green colours for HA-PB2, HA-PA, MP-NA and MP-NS cells in the matrix indicate that all these pairs tended to be inherited together from the same parental genome during reassortment.

3.3.4 Frequency and patterns of reassortment in Influenza C viruses

Screening of reassortment within 97 concatenated whole or near whole genome sequences isolated from Influenza C viruses revealed eight well supported putative reassortment events. These reassortment events predominantly involved transfers of the HE segment (5/8 events; Table 3.4) and occurred between parental sequences that had genetic distances ranging from 0.025 to 0.099 (Figure 3.5). The recombination region count matrix indicated that the HE segment can be exchanged between lineages with divergent origins.

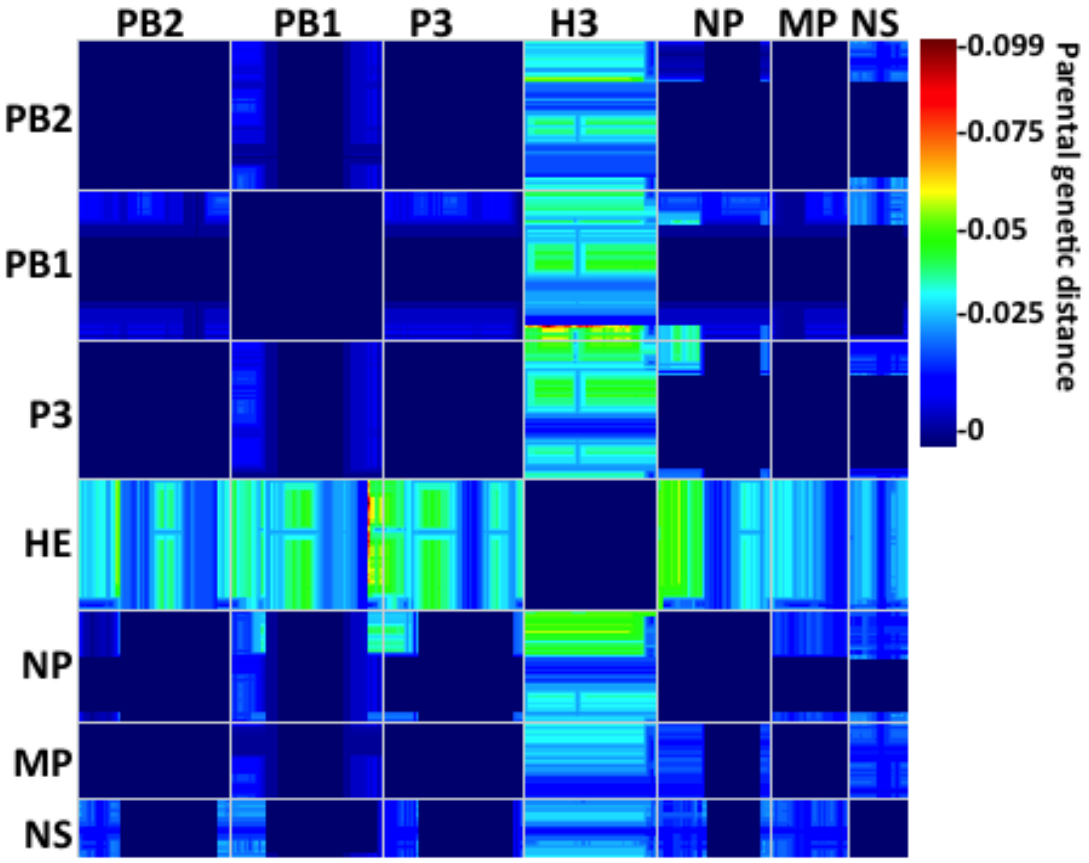


Figure 3.6 **Modularity matrix indicating local Influenza C virus parental sequence genetic distances across the entire genome:** The maximum genetic distances of parental viruses that yielded detectable reassortants/recombinants are depicted here. For any segment the graph should be read either from top to bottom or from left to right. Whereas more bluer colours indicate that reassortment events involved transfers of segments into genomes that had low genetic distances (calculated as unweighted Hamming distances) to the genomes from which the segments originated (i.e. the donor and recipient genomes were very similar in blue-coloured genome regions), redder colours indicate that reassortment events involved transfers of segments into genomes that had higher genetic distances to the genomes from which they originated (i.e. the donor and recipient genomes were genetically dissimilar in these genome regions). In reassortment events involving the HE segments, for example, parental virus's genome sequences tended to have genetic distances of up to 0.099 in the HE

segments which, in the case of the HE segments, had a distance across its entire length of >0.025 between the two parents in at least one pair of parental genomes that exchanged an NA segment.

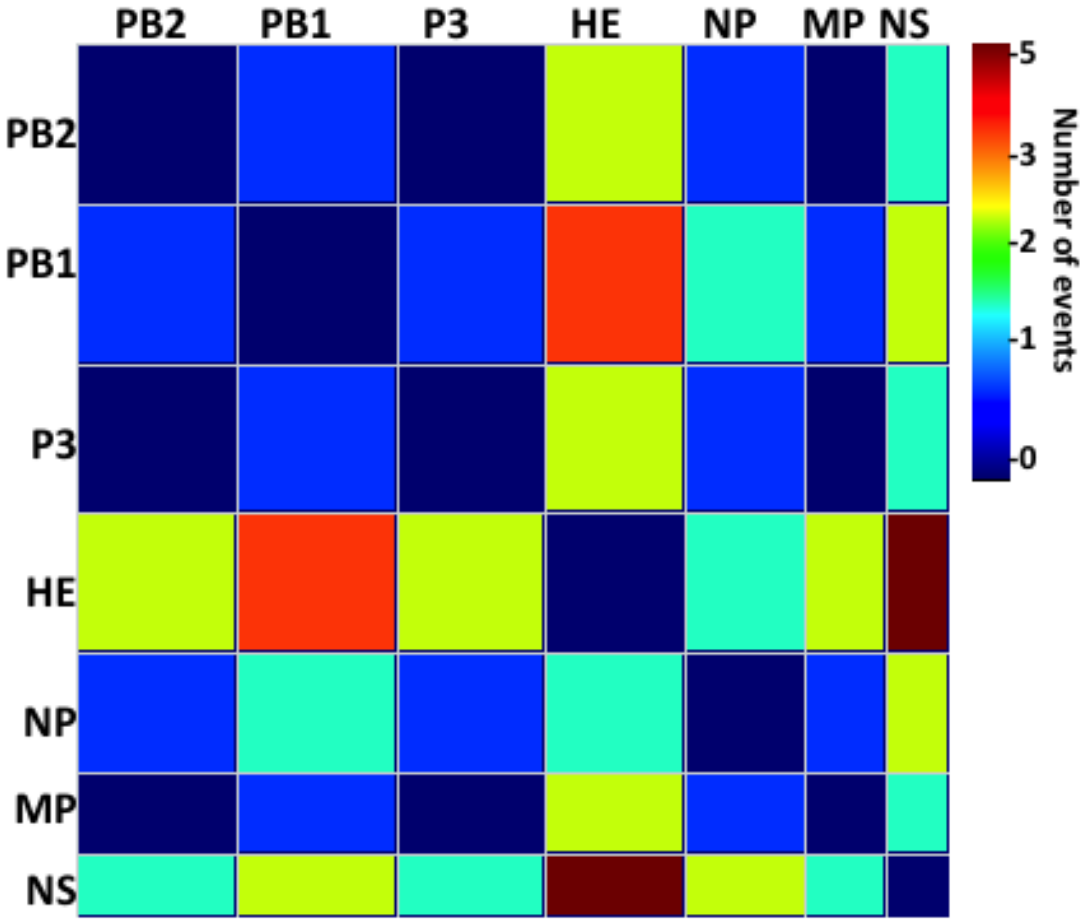


Figure 3.7 **Predicted regional counts of detected reassortment and recombination events** distributed through the genomic segments of the Influenza C virus. For each segment this plot should be read either from top to bottom or from left to right. With the HE segment row, for example, the yellow/orange/red colors for all segments indicate that whereas reassortment events involving the transfer of HE tend to frequently separate HE from the remainder of segments in the genome (i.e. HE tends to be transferred alone), the frequency with which these exchanges involve the separation of cognate HE and NS segments from one another is particularly high (as is indicated by the red color associated with HE-NS cells in this matrix), similarly the same is observed but to a lesser degree between HE-PB1. Conversely the blue-green color for the rest of cells in the matrix indicates that all these pairs tended to be inherited together from the same parental genome during reassortment.

3.3.5 ESTIMATING THE TIME SCALE (WHEN) AND SPATIAL ORIGINS (WHERE) OF REASSORTMENT EVENTS IN INFLUENZA A, B AND C VIRUSES BETWEEN 1927 - 2014

Towards understanding the time scale and geographical regions of reassortment events within the three types of Influenza viruses, I performed molecular clock and phylogeographic analyses employing coalescent Bayesian skyline and a discrete Bayesian phylogeographic model based analyses implemented within the BEAST package (as described previously). Given the computational intensity of these analyses, the whole genome dataset could not be used in their entirety; hence necessitating down-sampling to ensure much more computationally feasible analyses (see the methods section for details on how this was achieved). Subsampled datasets were obtained from the large whole genome sequence datasets used for reassortment analyses (also as described in the previous section). The discrete phylogeographic model enabled the reconstruction of phylogenies for three datasets i.e. Influenza A (315 sequences), Influenza B (200 sequences), and Influenza C (97 sequences) datasets.

Table 3.2 Timescale and geographical location of reassortment events in Influenza A viruses

Reassortment Even	Time of reassortmen	95% HPD	Location	Location probabilit
1	1917.53-1929.98	1900.27-1932.52	North America	0.32
2	1917.53-1929.98	1900.27-1932.52	North America	0.32
3	1921.52-1969.74	1917.53-1982.79	North America	0.31
4	1932.52-1969.74	1917.53-1982.79	North America	0.31
5	1932.52-1969.74	1917.53-1982.79	North America	0.31
6	1932.52-1933.72	1917.53-1946.66	North America	0.30
7	1932.52-1933.72	1917.53-1946.66	North America	0.30
8	1932.52-1933.72	1917.53-1946.66	North America	0.30
9	1932.52-1933.72	1917.53-1946.66	North America	0.30
10	1917.53-1929.98	1900.27-1932.52	North America	0.32
11	1932.52-1933.72	1917.53-1946.66	North America	0.30
12	1932.52-1933.72	1917.53-1946.66	North America	0.30
13	1932.52-1933.72	1917.53-1946.66	North America	0.30
14	1946.66-1951.53	1933.72-1956.66	North America	0.31
15	1965.67-1967.81	1962.41-1971.37	Europe	0.75
16	1932.52-1969.74	1917.53-1982.79	Oceania	0.47
17	1946.66-1952.53	1933.72-1956.34	Oceania	0.46
18	1946.54-1962.41	1927.47-1965.67	North America	0.44
19	1946.66-1952.53	1933.72-1956.34	Oceania	0.46
20	1946.66-1952.53	1933.72-1956.34	Oceania	0.46
21	1932.52-1933.72	1917.53-1946.66	North America	0.30
22	1973.94-1976.28	1964.64-1983.83	North America	0.67
23	1932.52-1933.72	1917.53-1946.66	North America	0.31
24	1946.96-1964.64	1933.72-1973.94	North America	0.31
25	1946.66-1964.64	1933.72-1973.94	North America	0.31
26	1994.31-1995.36	1993.98-1995.85	Asia	0.95
27	1977.98-1933.09	1964.27-1997.66	Africa	0.82
28	1994.31-1995.36	1964.27-1997.66	Asia	0.88
29	1971.49-1977.98	1964.64-1987.71	Africa	0.82
30	1932.52-1933.72	1917.53-1946.66	North America	0.31
31	1977.98=1993.09	1964.27-1997.66	Africa	0.82
32	1946.66-1964.64	1933.72-1973.94	North America	0.31
33	1932.52-1933.72	1917.53-1946.66	North America	0.31
34	1956.66-1964.27	1951.53-1977.98	North America	0.31
35	1932.52-1933.72	1917.53-1946.66	North America	0.31
36	1932.52-1933.72	1917.53-1946.66	North America	0.31
37	1946.66-1964.64	1933.72-1973.94	North America	0.31
38	1946.66-1964.64	1933.72-1973.94	North America	0.31
39	1946.36-1951.53	1933.72-1956.66	North America	0.31
40	1946.54-1962.41	1927.47-1965.67	North America	0.44
41	1973.94-1976.28	1964.64-1983.83	North America	0.67
42	1946.36-1951.53	1933.72-1956.66	North America	0.31
43	1983.83-1984.26	1976.28-1991.03	North America	0.86
44	1917.53-1927.58	1900.27-1932.52	North America	0.32
45	1932.52-1933.72	1917.53-1946.66	North America	0.31
46	1932.52-1933.72	1917.53-1946.66	North America	0.31
47	1932.52-1969.74	1917.53-1982.79	North America	0.31
48	2006.79-2007.98	2002.24-2008.43	North America	0.33
49	1946.36-1951.53	1933.72-1956.66	North America	0.31
50	1983.83-1984.26	1976.28-1991.03	North America	0.86

Table 3.3 Timescale and geographical location of reassortment events in Influenza B viruses

Reassortment Even	Time of reassortmen	95% HPD	Location	Location probabilit
1	2013.03-2013.30	2012.91-2013.40	North Americ	1.0
2	1984.75-1987.81	1980.12-1989.88	Asia	0.41
3	1984.75-1987.81	1980.12-1989.88	Asia	0.41
4	1990.39-1992.05	1990.15-1992.72	Asia	0.89
5	1984.75-1987.82	1980.10-1989.88	Asia	0.63
6	1984.75-1987.82	1980.10-1989.88	Asia	0.63
7	1989.88-1990.15	1984.75-1992.05	Asia	0.63
8	1989.88-1990.15	1984.75-1992.05	Asia	0.63
9	2000.63-2008.99	2000.08-2001.82	Oceania	0.99
10	1984.75-1987.81	1980.12-1989.88	Asia	0.41
11	2007.24-2008.99	2005.66-2010.87	Oceania	0.54
12	2001.80-2001.82	2000.63-2002.93	Oceania	0.96
13	2005.62-2006.24	2004.64-2007.10	Oceania	0.86
14	2000.08-2001.03	1999.73-2000.24	Oceania	0.99
15	2000.63-2001.24	2008.08-2001.82	Oceania	0.99
16	2006.55-2007.10	2006.24-2007.30	Oceania	0.86
17	1990.39-1992.05	1990.15-1992.72	Asia	0.90
18	1984.75-1987.81	1980.12-1989.88	Asia	0.41
19	1980.10-1983.11	1975.00-1987.82	Asia	0.42
20	1980.10-1983.11	1975.00-1987.82	Asia	0.42
21	1981-98-1985.57	1979.98-1986.93	Europe	0.50
22	1980.10-1983.11	1975.00-1987.82	Asia	0.42
23	1984.75-1987.81	1980.12-1989.88	Asia	0.41
24	1980.10-1983.11	1975.00-1987.82	Asia	0.42
25	1984.75-1987.81	1980.12-1989.88	Asia	0.41
26	2002.30-2002.93	2001.80-2003.29	Oceania	0.85
27	2005.66-2005.73	20003.01-2007.24	Asia	0.53
28	2005.66-2005.73	20003.01-2007.24	Asia	0.53
29	1989.88-1990.15	1984.75-1992.05	Asia	0.63
30	1984.75-1987.81	1980.12-1989.88	Asia	0.41
31	1984.75-1987.81	1980.12-1989.88	Asia	0.41
32	1984.75-1987.81	1980.12-1989.88	Asia	0.41
33	2000.63-2001.24	2000.08-2001.82	Oceania	0.96
34	1980.10-1983.11	1975.00-1987.82	Asia	0.42
35	1984.75-1987.81	1980.12-1989.88	Asia	0.41
36	2000.63-2008.99	2000.08-2001.82	Oceania	0.99
37	1984.75-1987.81	1980.12-1989.88	Asia	0.41
38	1990.84-1992.05	1990.25-1993.75	Asia	0.90
39	1989.88-1990.15	1984.75-1992.05	Asia	0.53
40	1990.84-1992.05	1990.25-1993.75	Asia	0.9
41	1980.10-1983.11	1975.00-1987.82	Asia	0.42
42	1989.88-1990.15	1984.75-1992.05	Asia	0.63
43	1980.10-1983.11	1975.00-1987.82	Asia	0.42
44	2005.62-2006.24	2004.60-2007.10	Oceania	0.86
45	1990.39-1992.05	1990.15-1992.72	Asia	0.63
46	1980.10-1983.11	1975.00-1987.82	Asia	0.42
47	1989.88-1990.15	1984.75-1992.05	Asia	0.63
48	1980.10-1983.11	1975.00-1987.82	Asia	0.42
49	1984.75-1987.82	1980.10-1989.88	Asia	0.41
50	1980.39-1982.05	1980.15-1982.75	Asia	0.89

Table 3.4 Timescale and geographical location of reassortment events in Influenza C viruses

Reassortment event	Time of reassortment	95% HPD	Location	Location probability
1	1937.54-1957.46	1920.45-1967.99	Japan	1.00
2	1932.61-1935.78	1916.45-1951.97	Japan	1.00
3	1935.78-1956.18	1919.90-1964.22	Japan	1.00
4	1974.29-1987.78	1964.22-1919.90	Japan	1.00
5	1948.75-1970.18	1946.93-1976.69	Japan	0.85
6	1932.61-1937.54	1916.45-1954.89	Japan	0.99
7	1923.84-1932.61	1906.36-1947.81	Japan	1.00
8	1923.84-1932.61	1906.36-1947.81	Japan	1.00

The analyses performed to explore the timelines of reassortment using concatenated genomes of the Influenza viruses suggest that reassortment is as on-going process as the inferred times of reassortment span the pre-sampling and sampling period of the sequences analysed Table 3.2-Table 3.4).

The Influenza A virus sequences analysed suggest that reassortant lineages were seeded into the various hosts as early as the 19th century. Further examination of the individual sequences involved indicate that avian strains of Influenza A underwent frequent reassortment compared to strains that infect other hosts (Appendix 6). The MCC tree recovered from analysis of Influenza A virus suggested that some of the earliest reassortment events likely happened in the early 20th century between 1917 and 1929 (1900-1932 (95% HPD)) in North America. Although most of the reassortment events detected had inferred geographical locations where they occurred to be North America (pp 0.31), other locations were also inferred as likely ancestral locations of reassortant lineages e.g. Europe (pp 0.23), Asia (pp 0.21) and Oceania (pp 0.21) Table 3.2; Figure 3.8).

Analysis of Influenza B virus sequences also suggested continued reassortment over time as reassortment events were detected spanning the entire spectrum of the sampling period (Table 3.3). The earliest reassortment events were inferred to have occurred mid 20th century

between 1980.10-1983.11 (1975.00-1987.82 95% HPD) in Asia whereas the latest events were estimated to have most likely happened in Australia between 2013.03-2013.30 (2012.91-2013.40 95% HPD) (Table 3.3, Figure 3.9). The MCC tree yielded from the posterior distribution of trees obtained from the BEAST analysis also revealed the MRCA of Influenza B virus's two major lineages likely existed in the mid 1970's to early-1980's i.e. 1979.28(1975.00-1983.11 95% HPD), an estimate that is consistent with previous studies attempting to infer the divergence time of the Influenza B virus Victoria and Yamagata lineages (Dudas *et al.*, 2015).

Spatiotemporal analysis of the reassortment of Influenza C virus indicated that the virus continues to evolve over time through reassortment. The earliest reassortment events from the sequences analysed herein suggest that they probably occurred between 1923 and 1932 (1906.36-1947.81 95% HPD) and the inferred location was Japan (Table 3.4, Figure 3.10).

The continued evolution of these Influenza viruses through reassortment warrants continuous surveillance in ecosystems as reasortment events has been associated the emergence of epidemics and pandemics as these events frequently enable the evasion of host immune responses, resistance to antivirals, and the crossing of host species barriers (Su *et al.*, 2015a; Joseph *et al.*, 2017).

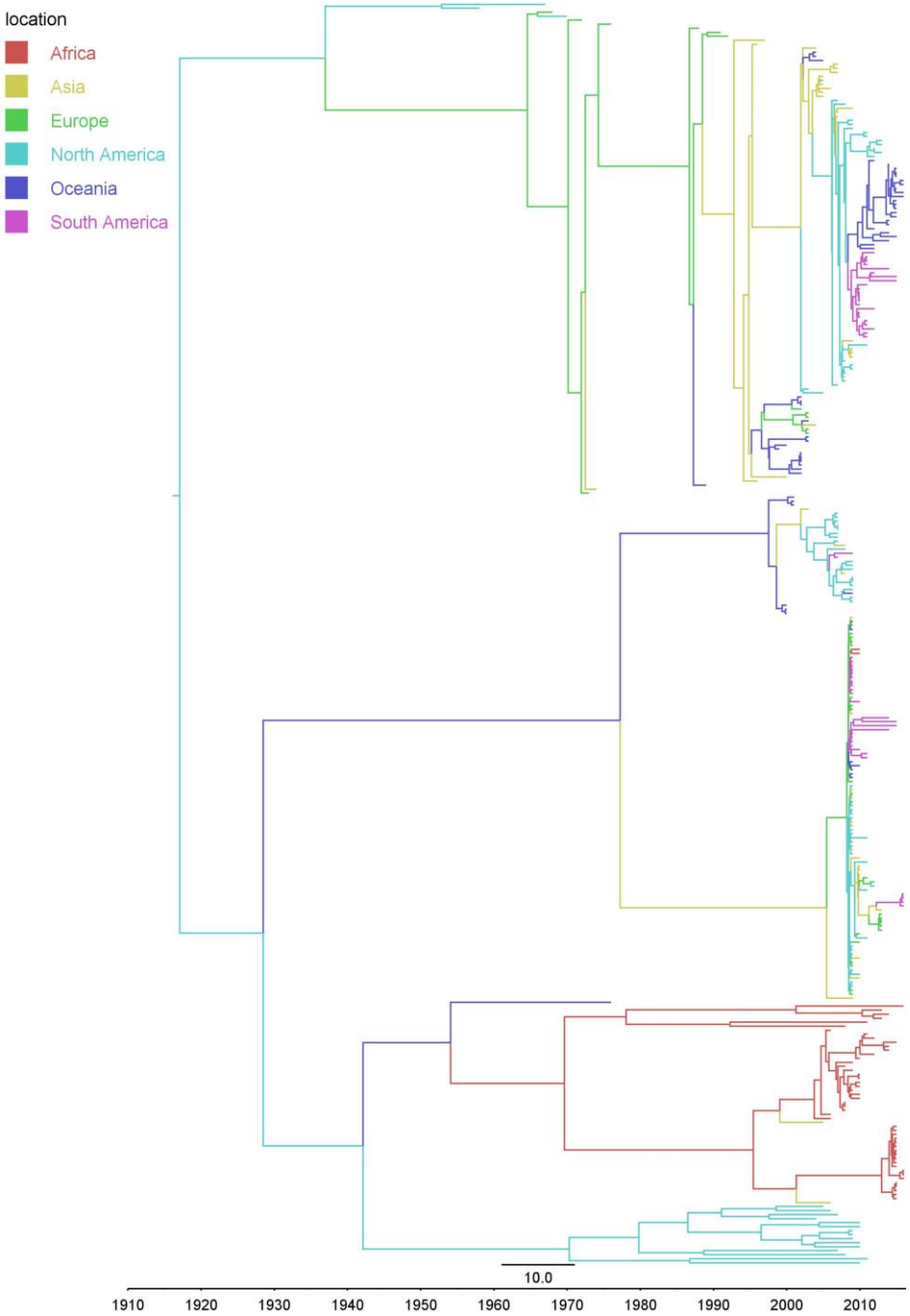


Figure 3.8 **Reconstructed MCC phylogeny illustrating tMRCA and ancestral geographical location of Influenza A virus ancestral lineages.** The ancestral locations of lineages that preceded the reassortant and pure viruses sampled between 1927-2014 were inferred . USA was inferred based on root state probability (Table 3.8) as the overall ancestral location of the Influenza Virus lineages that persisted to seed the present day reassortant lineages. The MRCA of all the present day lineages was estimated to have existed in 1820(95% CI: 1694-1837).

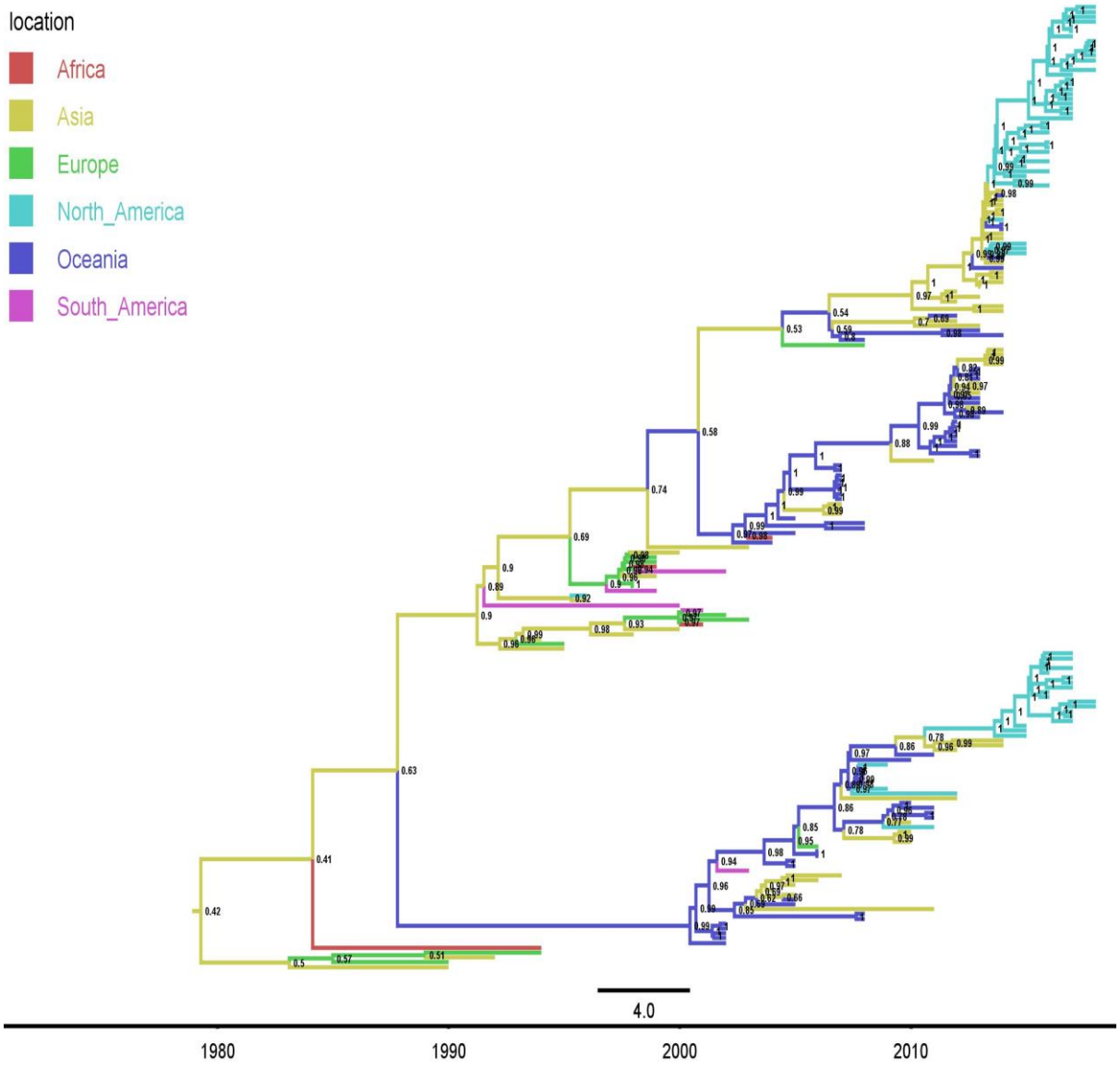


Figure 3.9 **Reconstructed MCC phylogeny illustrating tMRCA and ancestral geographical location of Influenza B virus ancestral lineages** illustrating ancestral sequences that seeded the currently circulating reassortant and pure lineages. The analysis indicates that the overall MRCA of these lineages existed in 1940 (1936-1941 95% HPD). It also provides additional inference of when the Influenza B virus lineages diverged into two major lineages: Victoria and Yamagata in early to mid 80's where the MRCA of the two lineages are estimated to have existed in 1977.28 (1973.00-1981.11 95% HPD) and the inferred ancestral location was in the Asia.

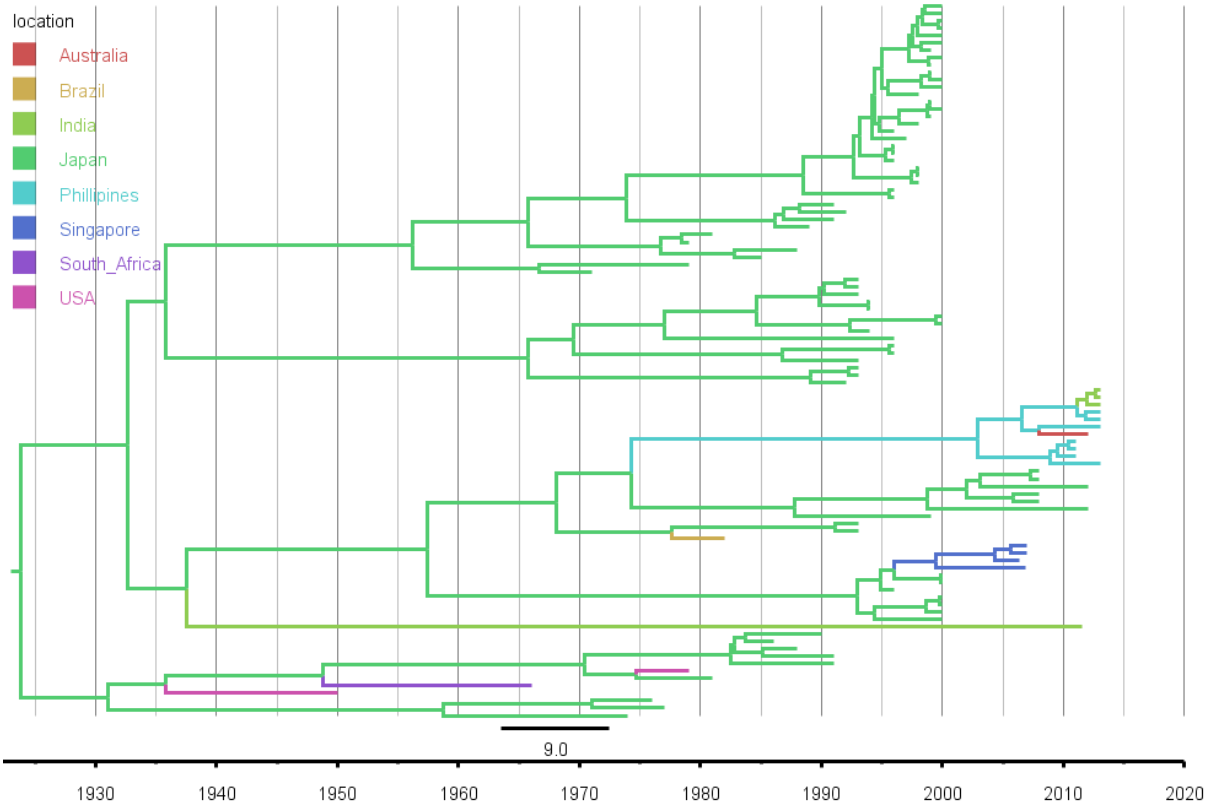


Figure 3.10 Reconstructed MCC phylogeny illustrating tMRCA and ancestral geographical location of Influenza C virus ancestral lineages to illustrate the timelines of when the MRCA of the predicted reassortant and pure lineages existed. The overall MRCA of Influenza C virus is inferred to have existed in 1923 (1906-1937 (95% HPD)) and the ancestral location is the Japan.

3.4 DISCUSSION

In this chapter, I performed a comparative analysis of the reassortment patterns of Influenza A, B and C viruses using whole genome data available in public sequence databases. The analysis I performed is the largest ever genomic sequence-based assessment of influenza virus reassortment and is one of the few that have attempted to investigate reassortment patterns across all influenza virus subtypes from a variety of hosts; other studies attempting to characterize influenza virus reassortment have generally foucsed either on individual subtype or on specific strain within particular hosts (Baker *et al.*, 2014; Lu, Lycett and Brown, 2014; Dudas *et al.*, 2015). I additionally,performed phylogeographic analyses to yield some insights into the geographical and temporal origins of the reassortant lineages that that were identified.

3.4.1 REASSORTMENT IN INFLUENZA A VIRUSES MOST FREQUENTLY INVOLVES TRANSFERS OF SEGMENTS ENCODING SURFACE PROTEINS

I identified 175 reassortment events within the analysed sample of Influenza A viruses that had been sampled over approximately eight decades (1927-2014). This analysis suggests that surface genes of Influenza viruses are transferred more frequently in reassortment events than are the internal genes of the Influenza A virus. Quantitative comparison of reassortment events further indicated that most events were identified within the HA and NA regions of the virus' genome. This could be attributed to the role of surface genes in driving Influenza infections in diverse hosts. Influenza A viruses have diverse host species and reassortment is a core mechanism required for crossing species barriers and to enhance continued spread among new hosts. This adaptive mechanism, commonly known as antigenic shifting, likely drives the observed degree of gene segment exchange among Influenza A viruses.

Inter-subtype reassortment seems to be particularly common between subtypes that naturally infect birds (such as H3N8, H1N1 and H5N1). Whereas reassortments amongst these viruses were observed that involve transfers of all the segments, HA and NA transfers were most common. This observed pattern of reasortment mirrors that observed previously in avian Influenza A viruses (Lu et al. 2014) and H3N2 sequences (Westgeest *et al.*, 2014).

Each of the eight Influenza B genome segments were inferred to have been transferred by reassortment in at least one instance. In contrast to Influenza A, there was no distinct difference between the internal and surface genes with respect to frequencies of transfers during reassortment (Figure 3.3 and Figure 3.4). The observation that the Influenza B components PB1, PB2 and HA tend to be co-ihereted with one another from the same parental virus during reassortent, is consistent with the previously surmised genetic linkage between these

components (Dudas *et. al.* in 2015). Given that the degree of diversity in Influenza B viruses is lower than that of Influenza A viruses is reflected in the fact that, relative to reassortment events in influenza A viruses, those occurring in Influenza B-viruses have tended to be between parental viruses that are more genetically similar. The divergence of Influenza B virus into two distinct lineages in the 1980's has also made possible the characterisation of reassortment lineages. The computational methods that were used here are only powerful enough to detect reassortmentment between viruses that are genetically quite distinct (i.e. sharing less than ~97% genome-wide sequence identity) and this may explain why the detected reassortment events were primarily between viruses belonging to the two main Influenza virus B lineages; i.e. our results do not imply that inter-lineage reassortment events are actually more common in nature than those occurring between viruses of the same lineage.

Influenza C virus is known to undergo reassortment but at a much lower frequency than that occurring in Influenza A and B species (Matsuzaki *et al.*, 2003). Although Influenza C has been detected worldwide most reported seasonal epidemics have been in Japan and the surrounding regions. Among the seven concatenated segments analysed, segment 4 which encodes haemagglutinin esterase (HE), was most frequently exchanged during reassortment events. Also, as with influenza A viruses, parental viruses of HE reassortants have tended to have been less genetically related to one another than parental viruses that exchanged other genome components. A recent longitudinal study on the genetic diversity and reassortment of Influenza C viruses spanning seven decades found several co-circulating Influenza C virus lineages in Japan (Matsuzaki et al., 2016). This analysis confirms the results of this other study which indicated that PB1, PB2, P3 and NP tend to be co-inherited from the same parental viruses during reassorthemnt events. It is possible that coinheretance of these internal genes offers a fitness advantage in that Influenza C viruses tend to maintains a constellation of co-inhereted internal genes from season to season.

3.5.2 Ancestral locations of parental genotypes/Reassortant genotypes are geographically dependent

A number of phylogeographic studies of virus dispersal patterns have revealed a host of factors that influence virus dispersal patterns (Dudas et al., 2015; Liang et al., 2010; Lu et al., 2014; Yoko Matsuzaki et al., 2016; Smith, Vijaykrishna, et al., 2009; Westgeest et al., 2014).In this chapter I have used similar phylogeographic tools to infer both where the Influenza A, B and C lineages arose, and when and where Influenza virus reassortants arose.

In these analyses the North American region was inferred to be the ancestral origin of Influenza A viruses (Table 3.2, Figure 3.8). This finding could, however, be strongly biased by the fact that the earliest efforts to track influenza virus epidemiology and characterize influenza virus genomic sequences were made in the USA. This sample bias could explain why my analysis was inconsistent with other studies that have indicated South Asia as the region where Influenza A first arose. Nevertheless, the USA as the origin of influenza A viruses supports a previous claim that the North America region may have been the 19th century source of the A/H1N1 strain that caused the 1918 Influenza pandemic (Worobey, G. Han and Rambaut, 2014).

My study also suggests that Influenza B viruses most likely originated in the Asia(pp 0.32) or Europe(pp 0.23) (Table 3.3; Figure 3.9). Although the current strains of Influenza B virus are presently dispersed throughout the world, the most severe Influenza B epidemics have occurred in the Oceania region. To date, however, there are few reports where Influenza B virus sequences sampled between 2007 and 2015 suggested the likely geographical origins of Influenza B virus and it therefore remains plausible that either the Asia/Europe or Oceania regions are the likely ancestral origin of these viruses (Dudas *et al.*, 2015; Langat *et al.*, 2017).
.

Although Influenza C virus still presents a potential health challenge globally, epidemics are presently localized in East Asia. Several lineages are currently co-circulating Japan (Speranskaya *et al.*, 2012) with ancient lineages having been isolated in 1947 and 2014 in North America (Peng *et al.*, 1994; Matsuzaki *et al.*, 2003, 2016; Speranskaya *et al.*, 2012). My analysis suggests that the HE and PE genomic regions have generally been most frequently transferred during reassortment events with the MRCA of these segments being Japan while the remaining segments are predicted to have a MRCA in the USA. As with the MRCA of the influenza A isolates examined here, the identification of the USA as the origin of these viruses (or at least a large portion of their genomes) is possibly impacted by the fact that most of the sequences analysed here (including the oldest sequences) originated in the USA.

3.6 CONCLUSION

Uncovering the reassortment patterns, geographical origins and movement dynamics of viruses such as Influenza A, B and C are key both to preventing future outbreaks and, when outbreaks do occur, to provide rational guidance on how these might be prevented from becoming global

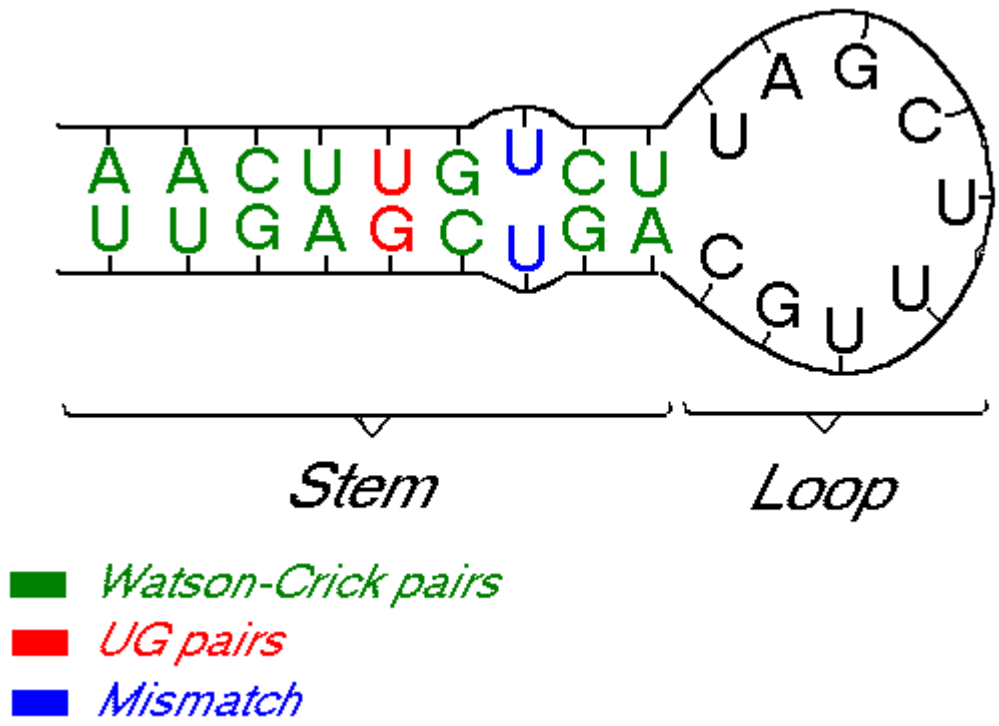
pandemics. It must be stressed that my inference of North America as the site where major influenza viruses have arisen and where these viruses have likely been most actively undergoing epidemiologically relevant reassortment events is potentially impacted by sampling biases. Given the tremendous global socio-economic impact of influenza viruses, however, I believe that such inferences should be taken seriously and that further efforts should be directed at both testing the hypothesis that North Amaerica is a major hotspot of epidemiologically important Influenza virus reassortment events, and determining the conditions in that region that may be contributing to the emergence of reassorted Influenza virus genomes.

CHAPTER 4

COMPUTATIONAL DETECTION OF RNA SECONDARY STRUCTURES AND EVALUATION OF THEIR IMPACT ON THE EVOLUTIONARY RATES OF ORTHOMYXOVIRUSES

4.1 INTRODUCTION

The architectures of single stranded RNA virus genomes can include a wide variety biologically functional nucleic acid secondary, tertiary and quaternary structural conformations (Simon and Gehrke, 2009). Such conformations arise primarily as a consequence of complementary base-pairing between the component nucleotides of the genomes and may play essential roles in the survival of RNA viruses (Gultyaev et al. 2010; Mathews et al. 2010; Moss et al. 2011; Dela-Moss et al. 2014; Priore et al. 2015). These roles may include the modulation of viral (Pedersen *et al.*, 2004; Kobayashi, Dadonaite, van Doremalen, *et al.*, 2016) or host (Liu *et al.*, 2016)(Iglesias and Gamarnik, 2014) gene expression, the determination of alternative splicing patterns (Dela-Moss, Moss and Turner, 2014b; Gultyaev *et al.*, 2016), the evasion of intra-cellular antiviral defenses (Guu *et al.*, 2008) and the physical stabilisation or compaction of genomes during packaging (Kobayashi, Dadonaite, Doremalen, *et al.*, 2016; Soszynska-Jozwiak *et al.*, 2017). Additionally, experimental study that applied thermodynamic energy minization and chemical mapping data of the RNP protein encoded by segment 5 (NP) predicted conserved structures at defined domains among Influenza A viruses, suggesting their biological importance (Soszynska-Jozwiak *et al.*, 2017).



Source: http://www.bioinf.man.ac.uk/resources/phase/manual/RNAMolecule.png

Figure 4.0 A RNA molecule secondary structure

Base pairing within secondary structures can take the form of canonical base pairs (i.e. Watson-Crick pairs of AU, UA CG and GC residues) and wobble pairs (like GU and UG) (Lorenz *et al.*, 2016). The "folded" RNA molecules that result as a consequence of these types of base-pairing can contain various types of RNA secondary structural elements including hairpin loops, pseudoknots, internal loops, multi-branch loops, and bulges (Anderson-Lee *et al.*, 2016).

Computational methods that attempt to predict RNA secondary structure from primary sequence data are commonly based on knowledge of the fundamental physical and chemical interactions that guide the secondary structure formation process (Mathews, 2005). One of the approaches used by these methods to predict the stability of base-pairing within potential secondary structures is the individual nearest neighbour-hydrogen bonding (INN-HB) model (Kierzek *et al.*, 2006). Using such a model to computationally compare the large numbers of RNA secondary structures that are possible for a given sequence remained challenging until the advent of new computational technologies as described in (Zuker, 2003; Markham & Zuker 2008). Currently, there are three approaches that are applied to RNA secondary structure discovery: thermodynamics approaches, sequence (molecular) evolutionary dynamics approaches and hybrid approaches that utilize both thermodynamics and molecular

evolutionary information (Hofacker, Fekete and Stadler, 2002; Freyhult, Moulton and Gardner, 2005; Bindewald and Shapiro, 2006; Bernhart *et al.*, 2008; Spirollari *et al.*, 2009; Washietl, Bernhart and Kellis, 2014) It is worth noting that each of these approaches have their own strengths and weaknesses and that none is clearly superior to the others under all circumstances (Seetin and Mathews, 2012).

In this chapter, I conducted the *in-silico* characterisation of the RNA folding patterns within the genomes of orthomyxoviruses. I additionally performed further association tests and molecular evolutionary analyses to determine whether predicted base-paired sites within folded regions tend to (i) be conserved (ii) co-evolve complementarily, (iii) be associated with synonymous substitutions that are lower than those associated with sites that are predicted to not base-pair and (iv) have detectably different nucleotide substitution rates to these non-base-paired sites.

Viruses within the family *Orthomyxoviridae* (hereafter referred to as orthomyxoviruses) have genomes with a number of characteristic features including that their genomes are negative sense RNA which is packaged as a single virion with distinct regions commonly referred to as segments which is, in-turn, are encapsulated within an envelope (Lamb et al. 2001; Falk et al. 1997; Leahy et al. 1997).

The nomenclature of orthomyxoviruses was derived from their conventional or 'straight' way of infecting epithelial cells and their affinity for the mucin proteins produced by these cells (Lamb, Krug and Knipe, 2001). Three of the seven known classes of orthomyxoviruses are the Influenza virus types A, B and C. The other four classes (sometimes also referred to as genera/species) include Isavirus (Falk *et al.*, 1997), Quaranja virus, Thogoto virus and the recently discovered Influenza D virus (Webster *et al.*, 1992).

Whereas Influenza A viruses infect a wide variety of avian and mammalian hosts (which, besides humans include pigs and horses), Influenza B and C viruses are generally only ever found infecting humans: although some reports have indicated that that they can sometimes also infect non-humans (Forrest and Webster, 2010). Isavirus is a fish virus that is known to infect salmon (Kibenge *et al.*, 2004), Quaranja virus infects arthropods and birds (Austin, 1978) and Thogoto virus is an arthropod virus that has been found infecting ticks and mosquitoes (Michael B. Leahy *et al.*, 1997).

Since orthomyxoviruses include some of the most important human pathogens, studying the potential impacts of genomic secondary structures on their evolutionary dynamics is certainly warranted (Szewczyk, Bienkowska-Szewczyk and Król, 2014). Such investigations could yield insights that foster the discovery of novel antiviral and/or vaccine targets that could be leveraged to combat the spread of these viruses.

It is important to point out, however, that the formation of secondary structures within a particular virus genome might not always have an impact on the biological functions of that genome. Identifying secondary structures that are suitable secondary-structure based drug or vaccine targets will depend largely on our ability to identify the secondary structural elements within a virus genome that are most likely to play a crucial biological role. Further, strong evidence exists that posttranslational gene regulation may rely on both the primary sequence of mRNA molecules, the secondary structures that these molecules fold into and the chemical modifications that are made to their bases (Lokody, 2014; Wang, Li and Gutenkunst, 2017).

The rationale behind the work presented in this chapter is that even if the secondary structures of orthomyxovirus mRNAs or genomic RNA molecules have only subtle impacts on the viability of these viruses, then these impacts could be readily detectable within the patterns of nucleotide variation that can be found within the hundreds of orthomyxovirus genome sequences that are presently deposited in public nucleotide sequence databases. By identifying patterns of nucleotide variation consistent with evolution favouring the maintenance of particular orthomyxovirus genomic secondary structural elements, I will in effect be able to identify those elements that are most probably functional.

4.2 MATERIALS AND METHODS

4.2.1 DATA PREPARATION

All available sequences of three orthomyxovirus viral species (4200,1800,150 whole genome sequences of Influenza A, B and C viruses respectively) were obtained from public databases. Influenza A, B and C virus sequences were obtained from NCBI Influenza Virus Resource: (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html), Influenza Research Database (http://www.fludb.org/brc/home.do?decorator=influenza) and Global Initiative on Sharing All Influenza Data (GISAID) at http://platform.gisaid.org/. These datasets were assembled between June and September 2013.

Using a similar approach, I retrieved all 394 available sequences of Infectious Salmon Anaemia Virus (Isavirus) and 48 Thogoto virus sequences from GenBank by running general search

queries in the web-based NCBI sequence database search interface. Given the segmented nature of orthomyxovirus genomes, some segments encode multiple genes. I prepared individual segment alignments for subsequent analyses. I aligned each segment separately using MUSCLE (Robert C Edgar, 2004) implemented in the SeaView program (Gouy, Guindon and Gascuel, 2010). The final datasets comprised 36 multiple sequence alignments that include eight individual segments of Influenza A, B and Isavirus, seven segments of Influenza C, and five segments of Thogoto virus. I used the program, RDP4 (Martin *et al.*, 2015), to select smaller datasets (consisting of 10 to 20 sequences each) that were representative of the diversity within each of these 36 datasets before performing the *in-silico* secondary structure predictions on these datasets. Several downstream analyses were performed on subsets of the final 36 datasets and hence a distinction was made to identify the initial overall datasets as being either being large (i.e. containing all available sequence data) or small (i.e. containing a representative selection of 10-20 sequences). The small datasets, which were used primarily for the secondary structure predictions, were selected from the corresponding large versions of the datasets by randomly selecting groups of sequences within these that shared 75% or less pairwise sequence identity (i.e. the selected sequences represented all the most divergent virus lineages).

4.2.2 IDENTIFICATION OF CONSERVED SECONDARY STRUCTURES WITHIN ORTHOMYXOVIRUSES GENOMES

I performed *in silico* prediction of conserved secondary structures within the genomes of the orthomyxoviruses using the Nucleic Acid Secondary Structure Predictor computer program available at http://web.cbio.uct.ac.za/~yves/nasp_portal.php (Semegni *et al.*, 2011). NASP employs a hybrid of thermodynamics and evolutionary analysis approaches implemented in the UNAFold program (Markham and Zuker, 2008). Prediction of the conserved folded regions in each set of aligned sequences was performed by first determining an ensemble of nearly minimum free energy (MFE) conformations for each individual sequence in the input alignment, and then identifying the subset of conserved secondary structures within the folded sequences that contributed most to their overall thermodynamic stability. This second step was achieved using a di-nucleotide shuffling based permutation test. For each of the input sequences in the smaller datasets, NASP generates sets of base-pairing matrices. These matrices are subsequently compressed into a consensus base-pairing matrix using a weighted sum of the pairing matrices for each of the input sequences. By using weighted matrices, the

technique can overcome unintended sampling bias in sequence datasets to ensure that similar structures in closely related sequences do not contribute disproportionately to the conservation scores generated for every structure that NASP identifies.

NASP further enabled the identification of statistically supported structures in the sequence sets – this subset of statistically supported structures for a specific dataset is herein referred to as the high confidence structure set (HCSS) of that dataset. Such HCSSs are characterised by their containing the ranked subsets of structural elements (from most conserved to least conserved) that collectively account for the predicted structures of the actual real sequences having significantly lower average minimum free energy estimates than those of 95% of the randomised dinucleotide-shuffled sequences. For the purposes of further characterizing a subset of predicted structural elements within the HCSSs that were most likely to be biologically functional, the top 20 structures in each HCSS were selected for further characterisation.

Specific parameters used for NASP runs included indicating that the sequences being folded were linear, and that folding should be inferred at 37^0 C under 0 magnesium and 1M sodium ionic conditions. HCSSs were determined by conducting 100 dinucleotide-shuffling permutations with a permutation P value cut off of 0.05. The overall output enabled the demarcation of conserved paired and unpaired sites within the analysed genomes.

4.2.3 TESTS OF SYNONYMOUS SUBSTITUTION RATES AT PAIRED/UNPAIRED SITES

Structured sites that tend to be in biologically functional regions of genomes are expected to remain relatively conserved compared to non-structured/non-functional sites (Wan *et al.*, 2011). Similarly, synonymous substitution rates at paired sites in structured regions are expected to be relatively higher than those at unpaired sites in unstructured regions (Belalov and Lukashev, 2013; Spielman and Wilke, 2015). Based on this, I set out to test the hypothesis that codons containing one or more base-paired sites have lower synonymous substitution rates than those containing only unpaired sites. This analysis was performed on nucleotide codon alignments of the 36 larger datasets representing individual orthomyxovirus genome segments of Influenza A, B and C, Isavirus and Thogoto virus. Each of the 36 alignments was split into codons that contained base-paired nucleotides at their third positions (called paired codon alignments) and codons that contained only unpaired sites (called unpaired codon alignments).

The paired and unpaired codon alignments were used to estimate synonymous substitution rates using the FUBAR method (Murrell *et al.*, 2013). FUBAR applies a time-reversible MG64 codon substitution model that uses a 61-by-61 codon substitution matrix. This method allows for independent distributions for non-synonymous and synonymous substitution rates and permits use of more rate classes to ensure subtle differences in selection pressures operating on individual codons can be discerned.

Analysis of selection pressures operating on codons in a given alignment using FUBAR relies on the use of an accurate phylogenetic tree. Given that recombination introduces a bias in phylogeny inference, the phylogenetic trees used as input for FUBAR analyses were generated from recombination free sequences obtained after screening and removing recombinant sequences. This was achieved by screening for recombinant sequences using the program, GARD (Kosakovsky Pond *et al.*, 2006) implemented in the HYPHY package (Kosakovsky Pond, Frost and Muse, 2005).

To determine whether paired codon sites have a significantly lower synonymous substitution rates than unpaired codon sites, I performed a Mann-Whitney U test on estimated synonymous substitution rates of codons in each of these categories for all of the 36 datasets. The p-values obtained from these tests were corrected for multiple testing using the step down method (Noble, 2009).

4.2.4 NEUTRALITY TESTS FOR PURIFYING SELECTION AT PAIRED SITES

I performed this analysis to test the hypothesis that paired sites within biologically functional secondary structures should be evolutionarily more conserved than unpaired sites in unstructured regions of the genome. Specifically, it is expected that paired sites should experience stronger selection disallowing nucleotide substitutions than unpaired sites. Tajima D and Fu and Li F tests were used to determine whether paired sites exhibit stronger evidence of purifying selection than unpaired sites (Tajima, 1989; Li, 1993). I estimated the Tajima's D and Fu & Li's F statistics for the paired and unpaired sites of the 36 datasets under study.

A permutation test was then applied to test whether the paired sites displayed significantly stronger evidence of negative selection than randomly subsampled unpaired sites: significance being inferred if the paired sites had D and/or F sores that were lower than 95% of D and/or F scores inferred for the randomly subsampled unpaired site datasets. Specifically, to obtain enough samples for the permutation test, 100 datasets were generated for each of the 36 large

codon alignments, each consisting of sites that were randomly sampled with replacement from the pool of unpaired sites. The Tajima's D and the Fu and Li's F statistics were computed for all the permuted datasets and the paired sites dataset. In all the 36 datasets, inference of stronger purifying selection at paired sites than at unpaired sites in the permutated datasets was estimated as being approximately equivalent to the proportion of times the D and F statistics computed for the paired site data set were lower than those for the 100 permuted datasets.

4.2.5 TESTING WHETHER PAIRED SITES COMPLEMENTARILY CO-EVOLVE

At paired sites in biologically functional secondary structures, it has been suggested (Jaeger, Turner and Zuker, 1989) that when mutations occur their impact will be minimised by compensatory mutations at the pairing partner sites that restore base-pairing. Using a customized version of the HYPHY's Spidermonkey coevolution script (Kosakovsky Pond et al. 2005; Pond et al. 2009; Muhire et al. 2014), I tested for evidence of complementary coevolution between paired sites within the 36 large datasets. The script compares the standard independent sites of a 4-by-4 HKY85 nucleotide substitution model to a 16-by-16 Muse-modified HKY85 (MG64) complementary coevolution model for any given pair of sites using a likelihood ratio test. This test relies on entries in the M95 16-by-16-substitution matrix that represent the changes that maintain base pairing and include both canonical and wobble base pairs (that are multiplied by a pairing factor, λ) and those involving changes between paired and unpaired sites (which are multiplied by $1/\lambda$)**.** Inference of complementary evolution was inferred for $\lambda > 1$, independent evolution (i.e. a standard HKY85 model was likely favoured) was inferred for $\lambda = 1$, and specifically non-complementary (or anti-complementary) coevolution was inferred for $\lambda < 1$.

As with any inference relying on phylogenetic trees, the analysis I performed could potentially be biased by recombination(Schierup and Hein, 2000). I screened all sequences in the 36 large datasets for mosaic sequences and removed the affected sequences.

4.2.6 TESTING FOR NUCLEOTIDE SUBSTITUTION RATES AT PAIRED/UNPAIRED SITES

I performed this analysis on separate paired and unpaired sub alignments of the 36 large datasets to test whether paired sites that are likely under strong purifying selection and have a tendency to be conserved have detectably lower nucleotide substitution rates than unpaired sites. I reconstructed time-scaled maximum clade credibility trees for each alignment using the generalised time reversible (GTR) model of nucleotide substitution and assuming four categories of gamma rate heterogeneity and coalescent constant population size models that

are implemented in BEASTv1.8.2 (Drummond *et al.*, 2012). Convergence of the analyses were visualized in Tracer (Rambaut, Drummond and Suchard, 2003) which was also used to assess whether MCMC had drawn enough samples (reflected in the ESS values >200) for all the parameters being estimated in the analysis. I generated the MCC phylogenies in TreeAnnotator (Drummond and Rambaut, 2007) and performed a comparative statistical analysis to assess any significant differences in the nucleotide substitution rates between paired and unpaired datasets.

4.3 RESULTS AND DISCUSSION

4.3.1 DATASETS

A total of 36 datasets consisting of individual segments of Influenza A, B, C, Isavirus and Thogoto viruses were collated from public nucleotide sequence repositories.

I retrieved Influenza A, B and C virus sequences from the NCBI Influenza virus resource database. Only strains with full genomes were included for Influenza viruses as limited whole genome sequences were available for Isavirus and Thogoto viruses sequences. I aligned individual segments for each virus separately using MUSCLE. For secondary structure detection ten sequences were selected from each alignment with a diversity threshold of 75% (i.e. all 10 sequences shared between 75% and ~95% identity with one another). The final alignments for each segment used in the subsequent analyses are as shown below (Table 4.1)

Table 4.1 Orthomyxoviruses datasets used to perform analyses in this study

Virus Type	Segment	Genes	No. of sequences
Influenza A virus	1	PB2	365
	2	PB1	407
	3	PA	470
	4	HA	425
	5	NP	332
	6	NA	411
	7	M1/M2	196
	8	NS1/NS2	208
Influenza B virus	1	PB2	120
	2	PB1	131
	3	PA	135
	4	HA	118
	5	NP	96
	6	NA	103
	7	MP	65
	8	NS1/NS2	58
Influenza C virus	1	PB2	96
	2	PB1	94
	3	P3	89
	4	HE	181
	5	NP	102
	6	M1/M2	128
	7	NS1/NS2	122
Isavirus	1		40
	2		48
	3		118
	4		508
	5		30
	6		118
	7		171
	8		24
THOV	2	GP	9
	3	M1/M2	4
	3	NP	9
	4	PA	9
	6	PB1	9

4.3.2 COMPUTATIONALLY PREDICTED SECONDARY STRUCTURES ARE DISTRIBUTED THROUGHOUT THE ORTHOMYXOVIRUSES GENOMES

The NASP analysis that was performed enabled the *in silico* detection of between 22 and 165 well supported secondary structures within the individual genome segments of the five analysed orthomyxovirus species (Influenza A virus, Influenza B virus, Influenza C virus, Isavirus and Thogoto virus).

Comparisons of homologous segments across each of the five species indicated the presence of a small number of structural elements that are potentially conserved at particular sites across all orthomyxoviruses. The subtle similarities in folding patterns between the homologous segments of different species suggests that some of these structures may be involved in important biological functions (Alexander P Gultyaev *et al.*, 2014; Muhire *et al.*, 2014; Kobayashi, Dadonaite, van Doremalen, *et al.*, 2016).

Apart from a minority of conserved structures, in general, there were clear differences in the density and distributions of predicted secondary structural elements both between homologous segments of different species and between the segments of each individual species (Appendix 7: Figures 4.1-4.8).

The NASP analysis suggests that within the analysed orthomyxovirus genomes, there potentially exist many more secondary structural elements than those which have currently been experimentally verified to have a biological function (Appendix 7: Figures 1-8).

4.3.3 CODON VERSUS NUCLEOTIDE LEVEL SELECTION AT PAIRED VERSUS UNPAIRED SITES

Biologically functional secondary structures occurring within coding regions are expected to be under purifying selection pressures that disfavour both changes to encoded protein amino acid sequences and changes to nucleotide sequences that disrupt the secondary structures. To test for evidence of whether codon sequences that coincided with detected secondary structures within the 36 orthomyxovirus large genome segment datasets might indeed be evolving under this type of "double" selection, I estimated the relative rates of synonymous substitutions at individual codon position within these datasets. Specifically, it was postulated that the synonymous substitution rates at codons that had a 3rd position that was predicted to be base paired within the predicted HCSSs should be significantly lower than those at codons containing only unpaired sites. I estimated synonymous substitution rates at individual codon sites using the FUBAR method (Murrell *et al.*, 2013).

Table 4.2: Comparative analysis of synonymous substitution rates at codon sites comprising of unpaired nucleotides and codons where the 3rd nucleotide position was predicted to be base-paired within HCSSs.

Dataset	Gene(s)	Median paired	Median unpaired	P – value
Influenza A 1	PB2	1.136	1.132	0.351
Influenza A 2	PB1	1.114	1.070	0.795
Influenza A 3	PA	1.165	1.203	0.016
Influenza A 4	HA	1.134	1.151	0.042
Influenza A 5	NP	1.041	1.128	0.003
Influenza A 6	NA	1.208	1.208	0.332
Influenza A 7	M1/M2	0.741	1.048	0.021
Influenza A 8	NS1/NS2	0.995	1.126	0.009
Influenza B 1	PB2	0.942	0.948	0.232
Influenza B 2	PB1	0.915	0.933	0.317
Influenza B 3	PA	0.987	0.998	0.178
Influenza B 4	HA	0.730	0.974	0.009

Influenza B 5	NP	0.893	1.005	0.183
Influenza B 7	M1/M2	1.194	0.969	0.689
Influenza B 8	NS1/NS2	0.828	0.744	0.701
Influenza C 1	PB2	0.953	0.994	0.001
Influenza C 2	PB1	1.039	0.987	0.349
Influenza C 3	PA	0.625	1.132	3.70e-07
Influenza C 4	HEF	0.879	0.895	0.221
Influenza C 5	NP	0.521	0.836	0.006
Influenza C 6	M1/M2	0.625	0.623	0.788
Influenza C 7	NS1/NS2	0.778	0.824	0.189
Isavirus-NS	NS1/NS2	1.128	0.872	0.999
Isavirus-PA	PA	1.0425	1.017	0.882
Isavirus-PB1	PB1	1.141	1.267	0.292
Isavirus-PB2	PB2	1.084	1.343	0.196
THOV-MP	M1/M2	4.001	5.801	0.282

After accounting for multiple testing in a Mann Whitney U test with a significance threshold of $p < 0.05$, I found that four of the Influenza A segments, one of the Influenza B segments and three of the Influenza C segments displayed significantly lower synonymous substitution rates at paired codon sites than they did at unpaired codon sites (Table 4.2). None of the Isavirus or THOV datasets displayed evidence of significantly decreased synonymous substitution rates at paired codon sites relative to unpaired codon sites. This suggests that, in some Influenza, A, B and C segments at least, the inferred secondary structural elements within the HCSSs are associated with additional constraints on the evolution of these segments and hence that at least a subset of the secondary structures identified in these segments are likely to be biologically functional.

It is, however; also, possible that the formation of secondary structures may have a direct impact on the underlying frequencies with which mutations arise within orthomyxovirus genomes. It is, for example, possible that base-pairing may protect nucleotides from chemical modification and subsequent mutation: a factor that might yield an observed decrease in

apparent synonymous substitution rates within structured regions without the need to invoke the action of "double" selection.

4.3.4 Additional evidence of stronger purifying selection at paired sites than at unpaired sites

I therefore performed an additional, more direct test, for differences in the strength of purifying selection at paired and unpaired sites. If base-pairing of sites within secondary structures that fall in coding regions is important for virus viability, it is expected that, whenever mutations arise at these sites, viruses carrying these mutations should have a lower frequency in the population than viruses that display mutations at unpaired sites. This type of "minor allele frequency" based evidence of purifying selection can be assessed statistically by neutrality tests such as those proposed by Tajima (Tajima, 1989) and Fu & Li (Li, 1993). Specifically, genomic regions under purifying selection are expected to yield lower values of both Tajima's D statistic and Fu & Li's F statistic than those evolving under neutral or positive selection (Tajima 1989; Li 1993). I therefore tested 20 of the 36 large datasets for evidence of mutations arising at base-paired within the HCSSs reaching lower frequencies in virus population than mutations arising at sites outside of the HCSSs.

Each alignment was split into two: One containing only sites predicted to be base-paired within the HCSSs and the other containing only sites predicted to be unpaired outside of the HCSSs. The Tajima D and Fu and Li F statistics determined for each base-paired alignment was then compared with those determined for permuted alignments sampled from the corresponding unpaired alignment as described earlier in methods section (Table 4.2).

The results suggest that for all but three of the analysed datasets paired sites within the HCSSs did indeed tend to have lower Tajima's D and Fu & Li's F statistics than those determined from unpaired sites (Table 4.3). For Influenza A virus, seven segments (1, 2, 3, 5, 6, 7 and 8) had lower Tajima's D and Fu & Li's F values at paired sites than at unpaired sites. The only exception was segment 4 that encodes HA and there is data that suggests it has the lowest structure conservation index (SCI) and lowest levels of synonymous substitution codon usage (SSCU) (Moss, Priore and Turner, 2011; Priore *et al.*, 2013; Dela-Moss, Moss and Turner, 2014a). The Influenza B virus analysis indicated that six of the segments had lower Tajima's D and Fu & Li's F statistics at paired sites than at unpaired sites. Again, the exception was segment 4. For influenza C five of the six tested segments had lower D and F statistics at paired sites than at unpaired sites, with the exceptional segment being 7. For Isavirus, the only dataset

analysed (for segment 2) also displayed lower D and F statistics at paired sites than at unpaired sites.

The permutation test that was used to test whether these trends in Tajima's D and Fu & Li's F statistics were significant revealed that in eight of the 17 cases where these statistics were lower for the paired sites than the unpaired sites, in less than 5% of the permuted datasets this trend was reversed (Table 4.3). For these eight datasets at least, there is statistically supported evidence of increased negative selection and not simply decreased mutation rates that are acting on paired sites and, therefore, that at least some of the predicted structures in these datasets are likely to be biologically functional. A significant increase in the negative selection in genes is a consequence of these sequences encoding biologically functional proteins. Among the proteins encoded by these segments are PA, HA, and NP in Influenza A viruses, HA in Influenza B viruses, and PB2, PA and NP in Influenza C viruses.

Paired sites in Segment 4 (encoding HA) of Influenza A and B viruses displayed higher values of the D and F statistics than did unpaired sites which indicates that the inferred structural elements within the HA gene are not detectably constraining the viability of viruses displaying mutations at sites that are predicted to be base paired. While this may mean that the potential structures that I have detected in this region are not biologically functional it could also be attributed to the selection on these structures being masked by opposing selective forces acting on the encoded protein. In this regard, HA is a surface protein that is evolving rapidly under positive selection in order to evade detection by host immune systems (Webster *et al.*, 1992).

Table 4.3: Tajima's D and Fu and Li F statistics for paired and unpaired alignments

Dataset	Tajima's D			Fu and Li's F		
Alignment	Paired	Unpaired	P-value	Paired	Unpaired	P-value
Paired Influenza A 1	0.146	0.463	0.05	1.495	1.611	0.05
paired Influenza A 2	0.166	0.465	0.21	1.444	1.504	0.33
paired Influenza A 3	0.296	0.816	0.04	1.478	1.652	0.02
paired Influenza A 4	0.075	-0.105	0.98	1.513	1.449	0.96
paired Influenza A 5	0.178	0.548	0.01	1.519	1.653	0.01
paired Influenza A 6	0.025	0.130	0.16	1.481	1.518	0.15
paired Influenza A 7	-0.034	0.167	0.23	1.354	1.457	0.14
paired Influenza A 8	0.010	0.414	0.001	1.468	1.614	0.001
paired Influenza B 1	-0.252	0.166	0.09	1.305	1.487	0.06
paired Influenza B 2	-0.109	0.073	0.36	1.302	1.328	0.45
paired Influenza B 4	0.435	0.226	0.65	1.295	1.196	0.68
paired Influenza B 5	-0.219	-0.361	0.69	1.231	1.218	0.52
paired Influenza B 8	-0.798	-0.380	0.09	1.001	1.182	0.02
paired Influenza C 1	-0.879	-0.849	0.43	1.082	1.103	0.33
paired Influenza C 2	-0.831	-0.772	0.43	1.021	0.973	0.66
paired Influenza C 4	-1.316	-1.055	0.001	0.973	1.085	0.001
paired Influenza C 5	-0.835	-0.176	0.09	0.448	0.944	0.02
paired Influenza C 6	-1.084	-0.894	0.23	0.683	0.811	0.12
paired Influenza C 7	-0.922	-1.089	0.89	1.068	1.007	0.89
paired Isavirus PB1	0.345	0.610	0.08	1.482	1.567	0.08

4.3.5 Evidence of complementary evolution at paired sites

One of the main selective forces that are expected to operate when arising mutations disrupt base-pairing within biologically functional secondary structures, is positive directional selection favouring the occurrence of compensatory mutations that restore base-pairing.

Specifically, in the predicted HCSSs of the various analyses orthomyxovirus genome segments, paired sites might be expected to, in some cases at least, display evidence of complementarily coevolving with one another. To test for evidence of this, I performed a statistical test on the paired and unpaired alignments and sites predicted to be coevolving versus those not coevolving using two-by-two contingency tests (Poon, Frost and Pond, 2009). Of the 36 analysed datasets five yielded statistically significant evidence of complementary coevolution between small numbers of base-paired sites within the HCSSs. These included the segment 4 (encoding HA) and segment 6 (encoding NA) datasets of influenza A virus, the segment 2 (encoding PB1) and segment 3 (encoding PA/P3) datasets of influenza C virus, and the segment 6 (encoding the F protein) dataset of Isavirus (Table 4.4).

Table 4.4 Chi square tests for coevolution at paired sites

Dataset	Chi square	P value
Influenza A 1	0.122	0.726
Influenza A 2	0.224	0.636
Influenza A 3	0.223	0.636
Influenza A 4	17.415	3.003e-05
Influenza A 5	0.107	0.743
Influenza A 6	4.808	0.028
Influenza A 8	1.942	0.163
Influenza B 1	0.001	0.993
Influenza B 2	0.051	0.822
Influenza B 4	0.015	0.903
Influenza B 5	0.219	0.639
Influenza B 7	0.189	0.663
Influenza B 8	0.249	0.617
Influenza C 1	1.647	0.199
Influenza C 2	5.537	0.018
Influenza C 3	7.177	0.007
Influenza C 4	0.619	0.431
Influenza C 5	0.003	0.956
Influenza C 6	0.166	0.683
Influenza C 7	1.383	0.239
Isavirus-F	726.918	2.2e-16

4.3.6 IMPACT OF SECONDARY STRUCTURE ON RATES OF NUCLEOTIDE SUBSTITUTIONS AT PAIRED AND UNPAIRED SITES

Given that secondary structures might influence both the rates at which mutations arise, and – if the structures are biologically functional – the rates at which mutations are either preserved

or purged by natural selection, it is conceivable that these structures might have an impact on substitution rates within orthomyxovirus genomes: impacts that could strongly influence the estimation of segment-wide substitution rates and confound attempts such as I have made elsewhere in this thesis to date past evolutionary events.

To determine the impact of secondary structure on the inference of substitution rates in orthomyxoviruses I applied the same Bayesian phylogenetic analysis approaches used in chapter 2 to infer the substitution rates of 31 out of the 36 segment datasets that had been previously split into paired and unpaired alignments. Given the decreased synonymous substitution rates inferred at paired sites previously, I anticipated that base-paired sites within the HCSS regions would have lower nucleotide substitution rates than those at unpaired sites outside the HCSS regions. BEAST, the computer program used for these analyses, allows estimation of absolute nucleotide substitution rates (as opposed to relative rates) by accommodating time-stamped sequence data as input and applying molecular clock models to the analysis of this data. Prior to analysis with BEAST, the best fitting nucleotide substitution models for each dataset were determined with the program, MEGA.

RNA viruses have some of the highest nucleotide substitution rates of all known organisms (Grenfell *et al.,* 2004; Drummond *et al.,* 2006; Biek *et al.,* 2015). The overall estimated rates of nucleotide substitution determined for paired site alignments were compared to those determined for the unpaired site alignments (Table 4.6). The general observed trend indicated that the sequences in the paired sites alignments had lower average nucleotide substitution rates than those in the unpaired sites alignments for all analysed segments of Influenza A, Influenza B, Influenza C viruses and Isavirus (Appendix 7 Figure 6). This analysis was not performed on THOV) because there were insufficient time-stamped genomic sequence data for these viruses.

Although the 95% credibility intervals of the nucleotide substitution rate estimates of matched paired and unpaired alignments were overlapping for all segments, the fact that nucleotide substitution rate estimates were consistently higher at unpaired sites than they were at paired sites strongly suggests both that at least some of the structural elements within the HCSS really do exist, and that the distributions of these structural elements have a direct impact on the rates at which different orthomyxovirus genome regions are evolving.

In accordance with other studies (Kühnert *et al.*, 2014; Worobey, G.-Z. Han and Rambaut, 2014), the highest nucleotide substitution rates were observed at the unpaired sites of segments encoding the HA and NA proteins (Table 4.6). This is likely at least partially due to positive

selection for immune evasion favouring the accumulation of non-synonymous substitutions within the coding regions of these segments.

Table 4.5 Nucleotide substitution rates of paired versus unpaired alignments within orthomyxoviruses.

Segment	Coding region	Nucleotide substitution rate	
		Paired	Unpaired
1	PB2	3.53×10^{-3}	$2.727 \times 10^{-3} \pm 0.577$
2	PB1	2.634×10^{-3}	$2.377 \times 10^{-3} \pm 0.55$
3	PA	3.227×10^{-3}	$3.289 \times 10^{-3} \pm 1.065$
4	HA	3.044×10^{-3}	$3.216 \times 10^{-3} \pm 0.0659$
5	NP	2.241×10^{-3}	$2.776 \times 10^{-3} \pm 0.735$
6	NA	3.568×10^{-3}	$3.264 \times 10^{-3} \pm 0.658$
7	M1/M2	3.839×10^{-3}	$2.478 \times 10^{-3} \pm 0.601$
8	NS1/NS2	6.626×10^{-6}	$2.708 \times 10^{-3} \pm 0.7$
Influenza B			
1	PB2	1.207×10^{-3}	$1.585 \times 10^{-3} \pm 0.544$
2	PB1	1.765×10^{-3}	$1.833 \times 10^{-3} \pm 0.597$
3	PA	1.487×10^{-3}	$1.845 \times 10^{-3} \pm 0.482$
4	HA	3.778×10^{-3}	$2.196 \times 10^{-3} \pm 0.488$
5	NP	1.749×10^{-3}	$1.584 \times 10^{-3} \pm 0.354$
6	NA		$1.965 \times 10^{-3} \pm 9.647 \times 10^{-2}$
7	M1/M2	3.666×10^{-3}	$1.892 \times 10^{-3} \pm 0.392$
8	NS1/NS2	1.303×10^{-3}	$1.511 \times 10^{-3} \pm 0.268$
INFLUENZA C			
1	PB2	2.832×10^{-3}	$1.435 \times 10^{-2} \pm 2.699$

2	PB1	1.887 $x10^{-3}$	1.86 $x10^{-3}$ ±1.938
3	P3	4.335 $x10^{-3}$	4.433 $x10^{-4}$±1.451
4	HE	6.655 $x10^{-3}$	6.578 $x10^{-4}$±0.721
5	NP	2.214 $x10^{-3}$	5.074 $x10^{-3}$ ±0.229
6	M1/M2	2.507 $x10^{-3}$	4.606 $x10^{-4}$ ±1.181
7	NS1/NS2	1.491 $x10^{-3}$	4.453 $x10^{-3}$ ±0.787
Isavirus			
1	PB2	2.824 $x10^{-3}$	1.809 $x10^{-4}$ ±0.704
2	PB1	1.058 $x10^{-3}$	1.76 $x10^{-4}$ ±1.05
3	PA	1.571 $x10^{-3}$	4.298 $x10^{-4}$ ±1.349
4	HA		4.043 $x10^{-4}$ ±1.415
5	NP		3.314 $x10^{-4}$ ±1.412
6	F		5.286 $x10^{-4}$ ±1.041
7	M1/M2		5.077 $x10^{-6}$±0.388
8	NS1/NS2	2.012 $x10^{-3}$	2.747 $x10^{-4}$ ±1.005

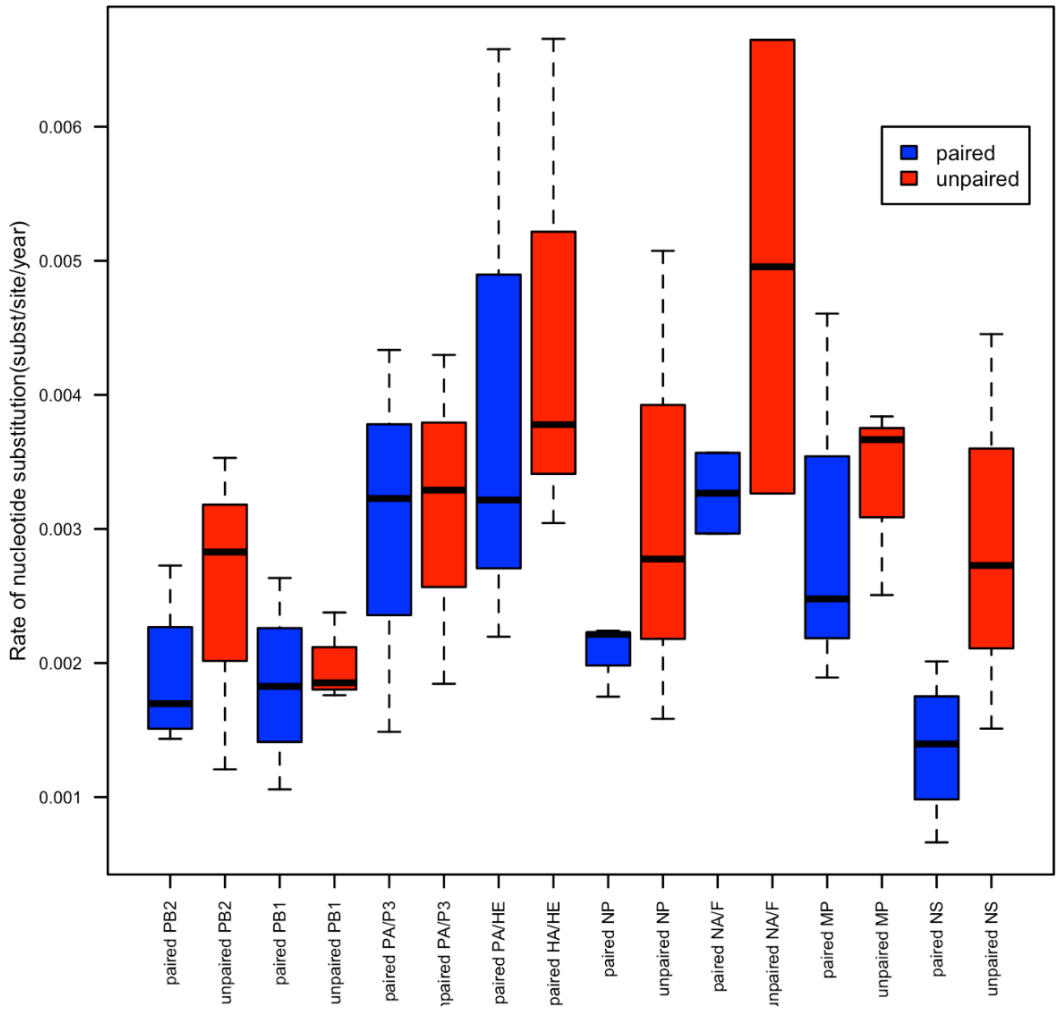


Figure 4.1: Boxplot illustrating average nucleotide substitution rates estimated in paired site (blue) and unpaired sites (red) alignments for the individual segments within genomes of various orthomyxoviruses. HA/HE and NA/F show the highest rates of nucleotide substitution. The box plots represent the 95% higher posterior density of the inferred nucleotide substation rates per site per year for paired and unpaired sites for each segment.

4.3.7 POTENTIALLY BIOLOGICALLY FUNCTIONAL CONSENSUS RANKED STRUCTURES WITHIN ORTHOMYXOVIRUS GENOMES

To better describe predicted structures and perhaps understand the roles that they may play in the survival strategies of these viruses, three criteria were used to rank structures in order of their likely biological functionality as applied in (Muhire *et al.*, 2014). The first was based on the conservation ranking of elements within the HCSSs that were produced by NASP. The second was based on the rate of synonymous substitutions within the base-paired sites of individual HCSS elements that fell within the coding regions. The third criterion was based on median probabilities of base-paired sites of individual HCSS elements coevolving with one another.

Additionally, the genomic locations of the highest ranking structural elements (i.e. those most likely to be biologically functional) in each of the species were carefully examined in the homologous region of the other species to determine the degree to which these elements have been conserved across all orthomyxoviruses. Although there were minor differences in the precise genomic coordinates of all the examined structures in the different orthomyxovirus genomes, various instances were encountered where distantly related genomes likely have highly conserved structural elements that are likely to have important functions in the biology of orthomyxoviruses.

PB2 Segment 1

With a size of approximately 2kb, segment 1 (PB2) is generally the largest orthomyxovirus segment. It plays an important role in the replication of these viruses (Webster *et al.*, 1992). Although the analysis I performed predicted 164, 165, 162 and 151 individual RNA secondary structural elements in the Influenza A, B, C and Isavirus segment sequences, respectively, the ranking based on NASP and other associated tests predicted one potential conserved structure along segment 1 of four species within 800-1000 base pair region in the 5'-3; upstream rgion (Figure 4.10). Due to limited availability of sequence data of the THOV, it could not be included in some of the segment specific assessments of the existence of conserved structures. Detected structures clustered within three regions along the PB2 segments. These regions were between 800-1000, 1030-1130 and 2000-2200 (Figure 4.10). It is also interesting that many of the structures in this gene occur near the 5` and 3` ends suggesting that they could perhaps play a role in the initiation or termination of virus transcription. Moss et al (2011) have previously described structured regions within the eight segments of Influenza viruses (Moss, Priore and

Turner, 2011). They predicted potentially conserved RNA secondary structures by employing an array of computational and *in vitro* methods that included a combination of amino acid and nucleotide sequence analyses and thermodynamics-based folding implemented in the RNAz program (Li *et al.*, 2010). In segment 1 (PB2), they found that the 5` and 3` ends of the PB2 gene are highly conserved and had partial complementarity to, and therefore likely participated in base pairing with, the promoter region. Many of the identified structures likely formed pseudoknots which are known to play a role in viral gene expression and replication. Here I also identified such structures in segment 1 of Influenza C virus.

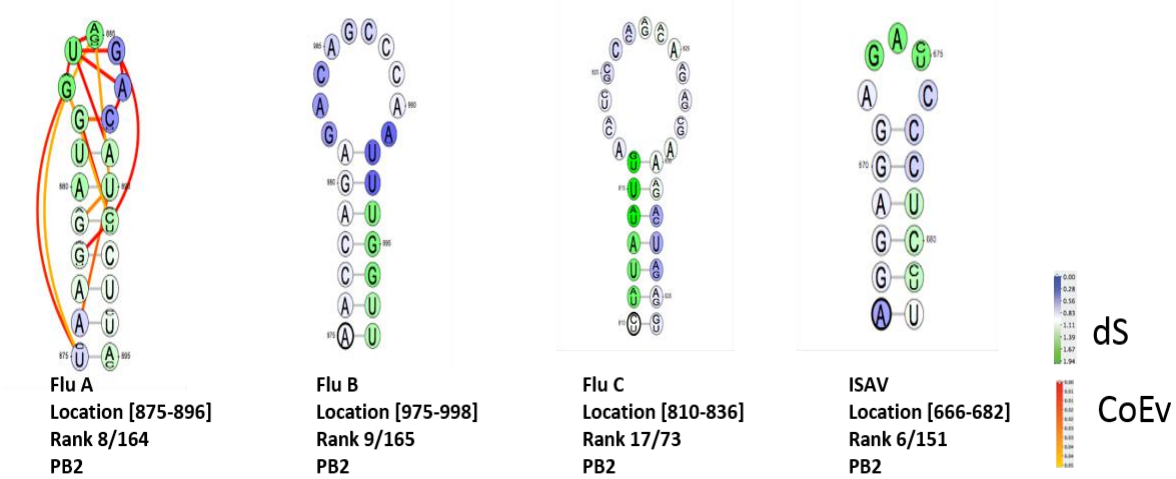


Figure 4.2 Predicted potential conserved structures within the Polymerase Basic 2 (PB2) i.e. segment 1 of the orthomyxoviruses. Potential conserved structures within genomes of Influenza A, Influenza B virus, Influenza C virus and Isavirus within the HCSSs of the various species were observed in the region: 800-1000. These region is associated with major splice sites within the PB2 gene. The red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method

PB1- Segment 2

Segment 2 is known to encode two alternative gene products, PB1-F2 and PB1-N40 (Priore *et al.*, 2015). Secondary structures have been previously detected using a combination of thermodynamics-based folding, sequence variation analysis and other hybrid methods at the splicing sites of the PB1 gene of the Influenza A, B and C viruses structures between positions 75-200 are thought to form part of the alternative splice initiation site of PB1-F2 and PB1-N40 gene products. In addition to these structures, the analyses I performed here also revealed other structures that are potentially conserved within the PB1 genes of all orthomyxoviruses (Moss, Priore and Turner, 2011; Priore, Moss and Turner, 2012, 2013; Dela-Moss, Moss and Turner, 2014b, 2014a; Priore *et al.*, 2015). Many of the conserved secondary structures of this segment were stem loop and bulge structures (Figure 4.11).

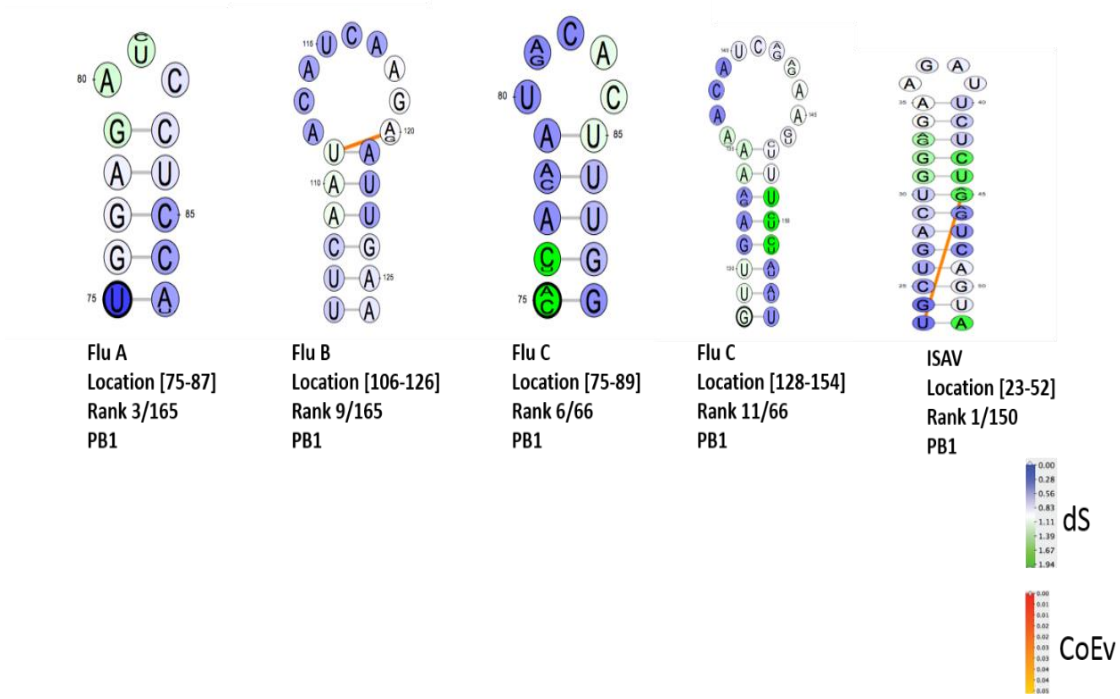


Figure 4.3 Putative conserved structures within three regions along segment 2 (PB1) of Influenza A, B, C, Isavirus viruses. Conserved stem loop structures were identified between positions 23-175 among the five species of the viruses. Similarly loop structure was detected in positions 93-180 in this segment in all species of the orthomyxoviruses. The red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method.

Segment 3 (PA/P3)

The Polymerase Acidic (PA) gene from segment 3 in Influenza A, B, Thogoto and Isavirus and P3 in Influenza C viruses forms part of the heterotrimeric polymerase complex in orthomyxoviruses, has endonuclease activity and is known to play a role in cap cleavage during replication. The analysis of the PA encoding segments of the five analysed orthomyxovirus species predicted mainly stem loop and bulge RNA secondary structures in three regions: 75-250 and 1600-1800 (Figure 4.12).

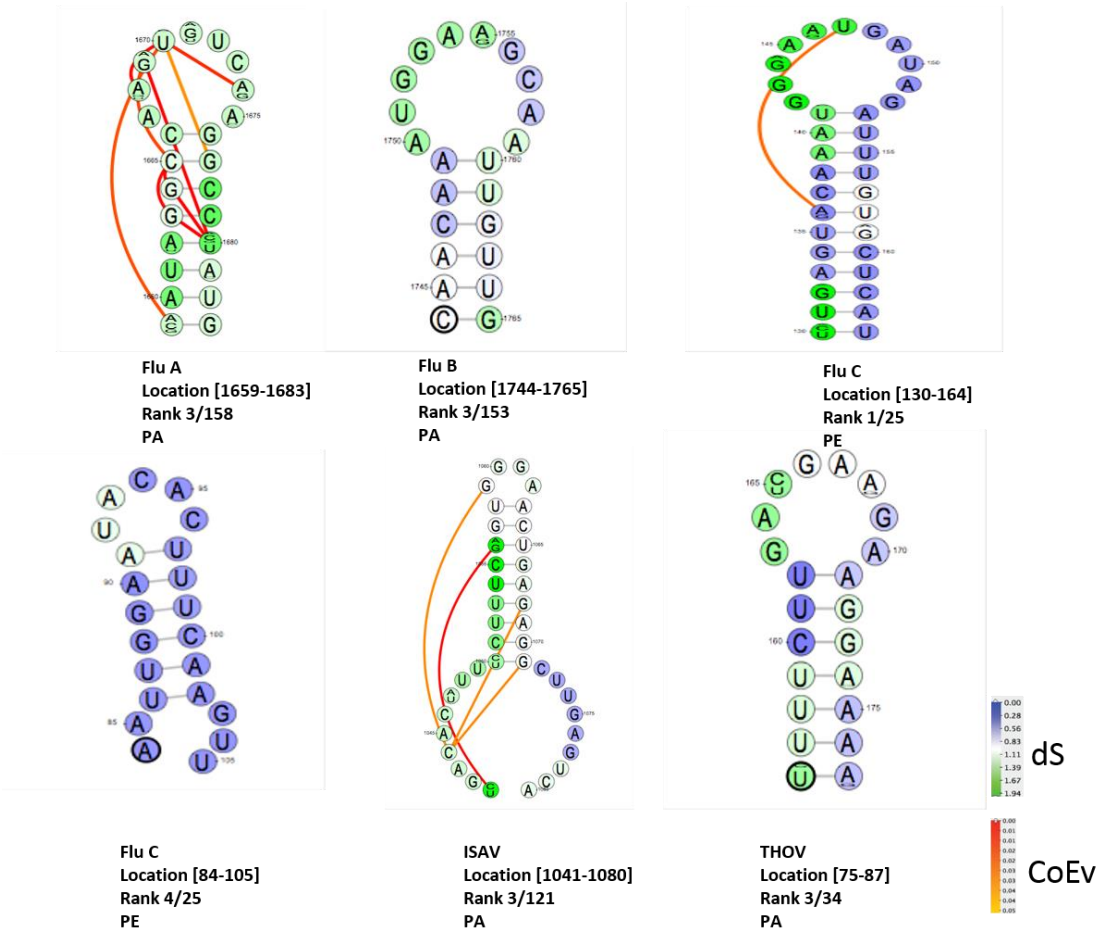


Figure 4.4 Predicted potential conserved RNA secondary structures within segment 3 (PA) of five species of orthomyxoviruses (Influenza A, Influenza B, Influenza C, Isavirus and Thogoto Virus). A stem-loop structures was predicted and likely showed to be conserved across all the fives species at positions 75-250 and 1600-1800. The red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method.

Segment 4 (HA/HE)

Segment 4 encodes the major surface glycoprotein of Influenza viruses, HA. The variant of HA found in Influenza C is referred to as the haemagglutinin esterase (HE) gene. High confidence and consensus ranked structure across all viruses in this *Orthomyxoviridae* family were mapped to three regions of segment 4: 900-1100 in Influenza A, B, C virus and between positions 200-400 in Influenza A and C virus (Figure 4.13). Similar to previous studies, these structures are predicted at the 5' splice sites of the various analysed orthomyxoviruses (Dela-Moss, Moss and Turner, 2014b).

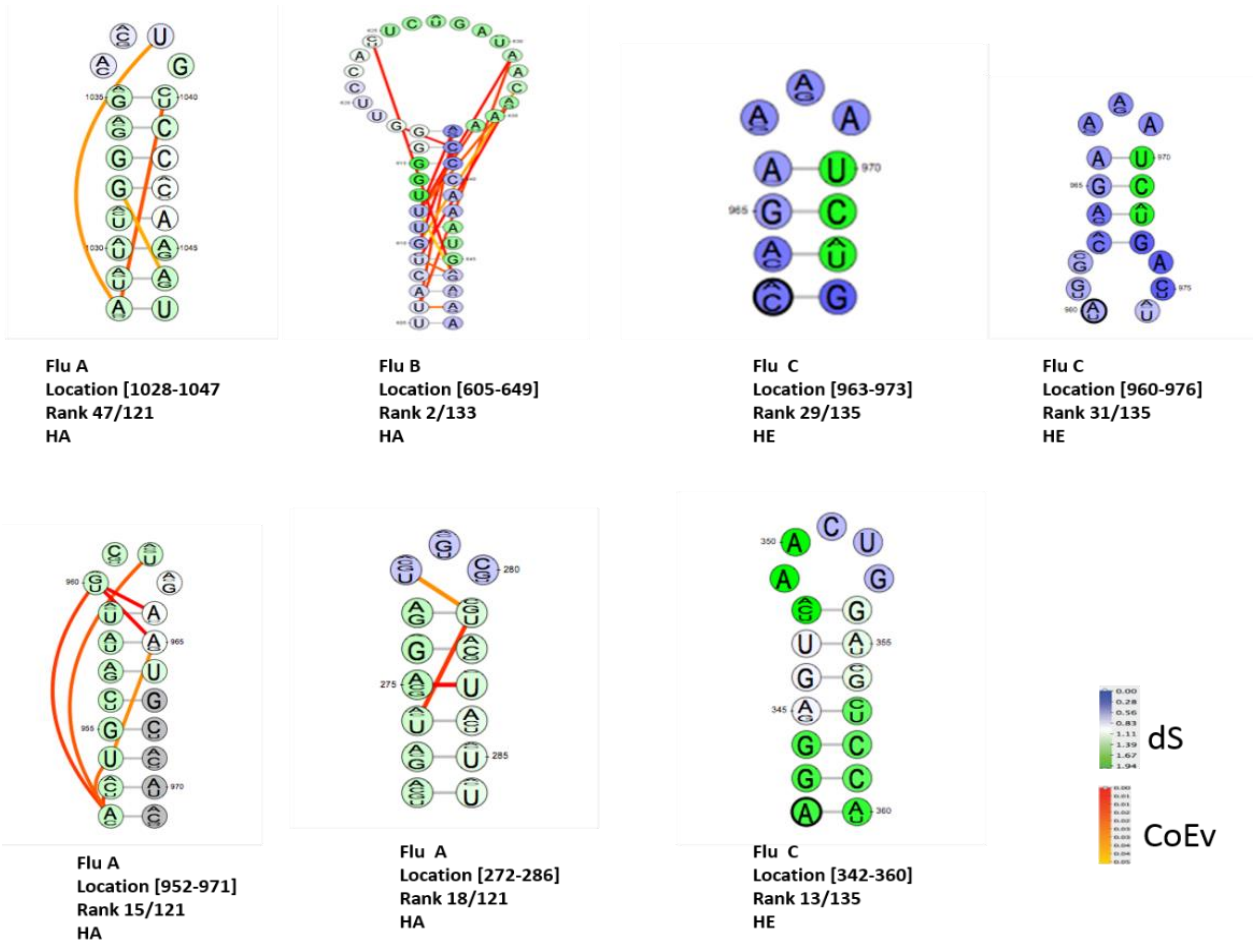


Figure 4.5 Segment 4 (HA/ HE) consensus ranked RNA structures common in three (Influenza A, B, and C species of orthomyxoviruses. No similar structures within the same positions were detected in Isavirus and THOV virus within this segment. The red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method.

Segment 5 (NP)

Segment five encodes a ribonucleoprotein in orthomyxoviruses and is thought to play an anchoring role for efficient packaging of all other segments into a viable virion. The consensus ranked structures are mostly stem loop and bulge structures (Figure 4.6). The structured regions cluster at splice site junctions at three regions 30-140, 1000-1100 and 1400-1500 along the segments. This is consistent with two previous studies that reported folding in approximately the same regions (Moss et al. 2011; Dela-Moss et al. 2014b).

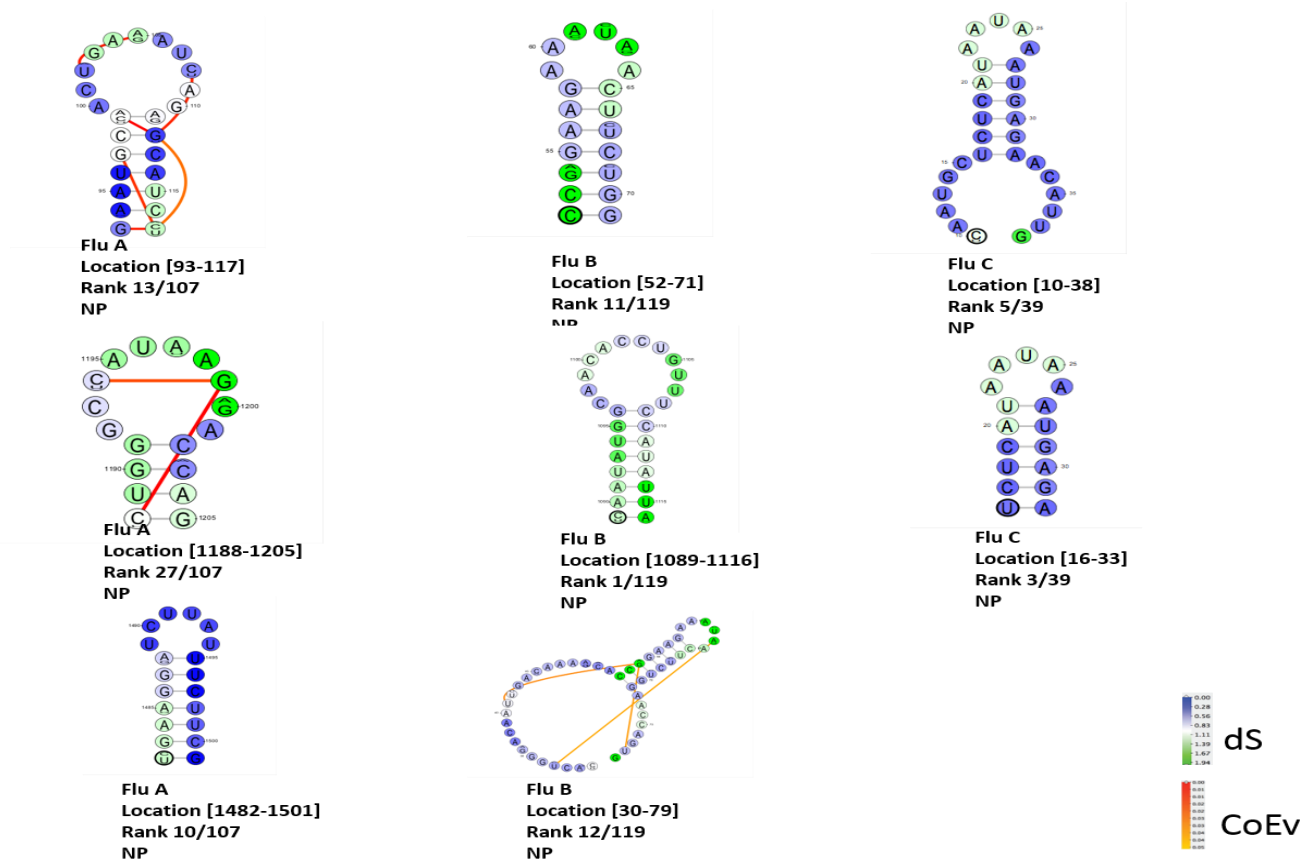


Figure 4.6 Putative potential conserved RNA secondary structures within NP gene (segment 5) in datasets representing Influenza A, B, C viruses. No similar structures were observed within the same positions from the analysis of the NP sequences of Isavirus and THOV virus. In this figure, the red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method.

Segment 6 (NA)

The neuraminidase forms part of the surface glycoproteins in Influenza A, B and Isavirus viruses. There was no sequence data on this segment for Thogoto virus and Influenza C virus lacks this segment hence no consensus structures are reported for these viruses for this analysis. NA is involved in the production of new virions by enhancing the budding of new virus and ensuring their release by cleaving a sialic acid residue of the host cell receptor. No conserved structures were detected across the five species, however, the few structures detected in Influenza A virus NA were mainly stem loop structure that ranked highly at region 102-129 in the 5' to 3' direction (Figure 4.7).

Figure 4.7: No conservation of RNA secondary structures was detected within the segment 6 (NA) of Influenza A, B and Isavirus viruses. The circular conformations detected along positions 500-650 across the segments of each species suggest limited or no conserved folding within segment 6 within this *Orthomyxoviridae* family. Sites predicted to be coevolving are shown by red-orange lines whereas the estimated rates of synonymous substitutions among sites are shown in blue-green colour.

Segment 7 – M1/M2 (Influenza C -segment 6)

Segment 7 encodes the matrix protein of Influenza A, B, C and THOV viruses while in Influenza C it is segment 6 that encodes the Matrix protein. The matrix protein transcript undergoes alternative splicing to form M1 and M2 components of the protein. Structure predictions within this segment and subsequent analyses indicate that it is among the most structured of the orthomyxovirus segments. The ranked consensus structures included a mix of multibranch stem loops, stem loops and bulge conformations (Figure 4.8). Structures mapped to 75-200, 250-400 and 700-900 regions are close to the major splice sites within the genes encoded for by this segment.

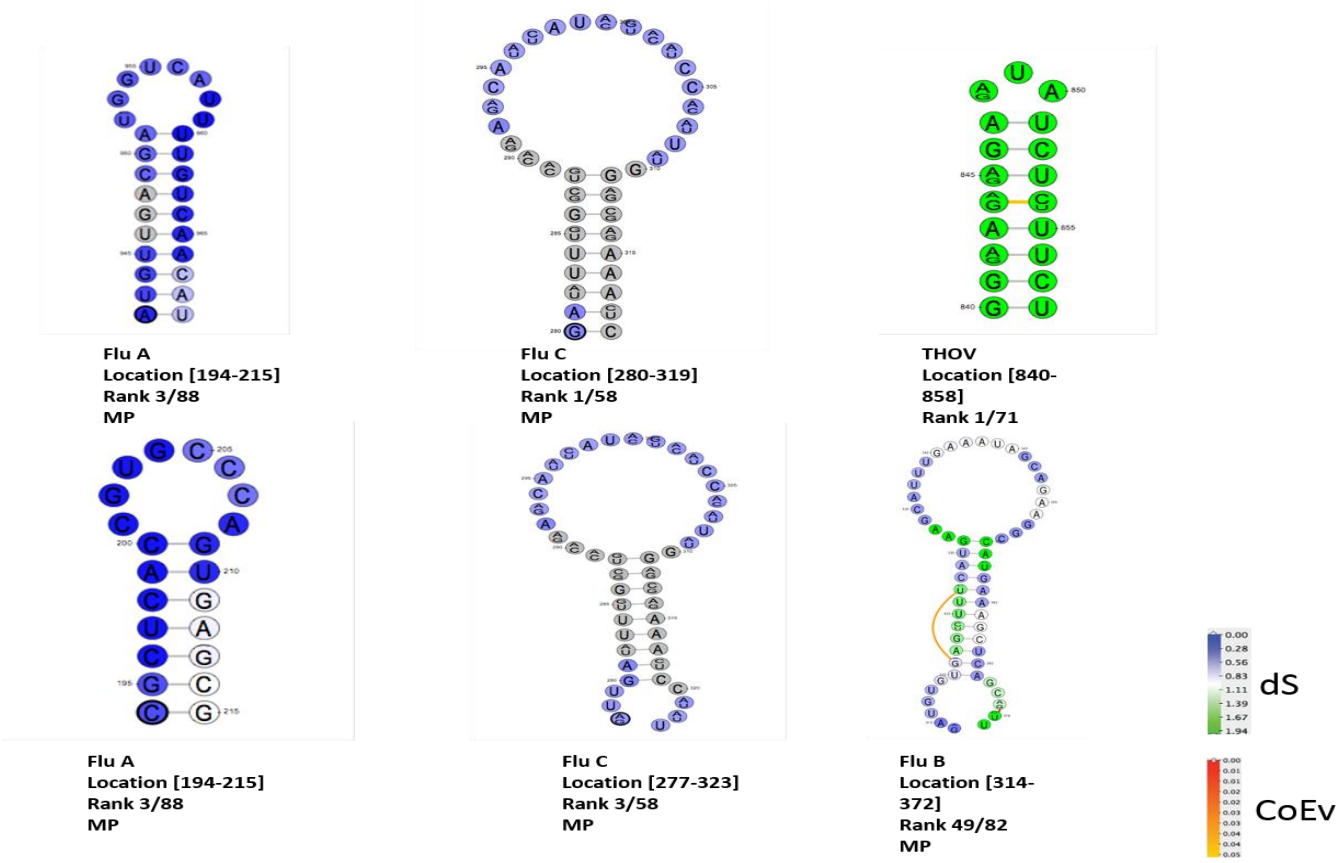


Figure 4.8 Predicted potentially conserved RNA secondary structures within the Matrix gene (M1/M2) that represents segment 7 in Influenza A, B, Isavirus and Thogoto virus and Segment 6 in Influenza C. The secondary structures within this segment were observed to be mainly stem and loop conformations. The red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method.

Segment 8 - NS1/NS2 /Segment 7(Influenza C)

The non-structural protein (NS1/NS2) is encoded by segment 7 in Influenza C and segment 8 in Influenza A, B and Isavirus. It has two components: NS1 and NS2. The predicted folding patterns of this segment indicate that it is generally the most structured segment within the analysed orthomyxovirus genomes. Multibranch stem loops were detected among the highly ranked structures in Influenza A virus. In other species, most of the consensus ranked structures were in the form of stem loops (Figure 4.9). Three structured regions were found between positions 30-150 in the NS1 sub-segment, and between 500-600 in the NS2 sub-segment. This suggests that alternative splice sites for the overlapping open reading frame of the two genes may be highly structured.

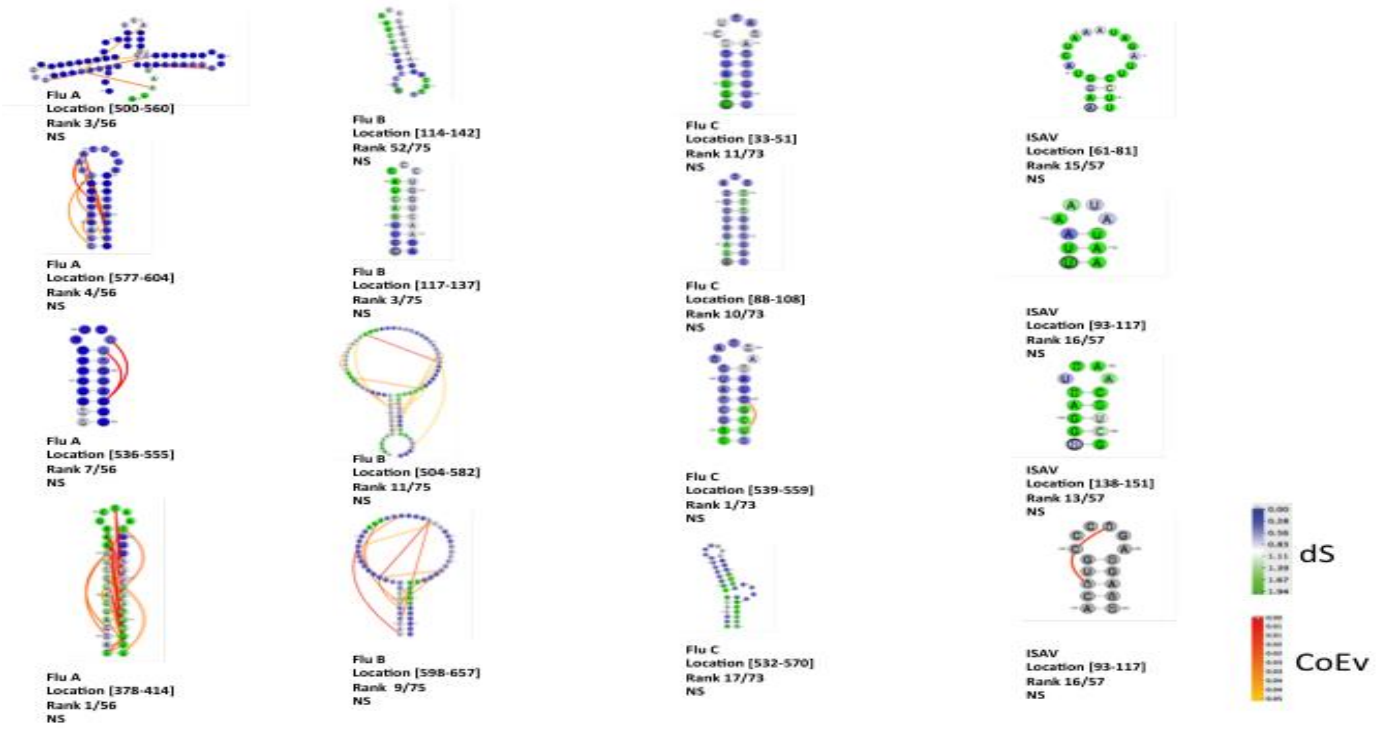


Figure 4.9 High confidence structure sets (HCSSs) within the Non-structural gene (NS1/NS2 (NEP)) commonly known us segment 8. Within Influenza A virus multi-branch stem loop structure showed that some conformation can be more complex than others. The red-orange-yellow lines depict sites predicted to be coevolving (CoEV) in the structured sites, whereas the green-blue coloured nucleotides show the estimated rate of synonymous nucleotide substitution rates (dS) estimated using FUBAR method.

4.4 CONCLUSION

Orthomyxoviruses are characterised by segmented RNA genomes that are co-packaged into a single enveloped virion. This study set out to firstly characterise the secondary structures of orthomyxovirus genomes and then evaluate the impact of these structures on the evolution of orthomyxoviruses. Using sequence data from five species I report herein, an unprecedented attempt to computationally discover and characterise RNA secondary structures within the genomes of these viruses. I further performed additional tests that provided additional layers of evidence that at least a subset of the identified structured regions has biological functions that have likely been preserved by selection since the origins of orthomyxoviruses.

This study shows that although individual segments have numerous unique structural elements, there are also several elements that appear to have been conserved across multiple orthomyxovirus species. In some cases, similar patterns of conserved folding have been observed in previous studies using different analytical approaches to those applied here. For instance, previous computational investigations reported structured regions within various Influenza A virus segments (Dela-Moss et al. 2014a; Moss et al. 2011). These studies suggested that segment 8 had more structured regions based on a combination of thermodynamics-based folding predictions and sequence analyses. The same observation was made here although my study generalized this previous observation by covering five species within the orthomyxovirus family.

Further, a study that focussed on the non-structural gene segment of Influenza A virus revealed conserved hairpin and pseudoknot structures by computational analysis within the NS1 and NS2 open reading frames at the at the 82-148 and 497-564 splicing sites that was particularly prominent in H5N1 sequences. Variations in the structure conformations in these regions in different orthomyxovirus species provides additional evidence that the evolution of structured regions within the NS segment may play a crucial role in host adaptation (Vasin *et al.*, 2016). Specifically, the variations in the folding patterns revealed in my study may be attributable to the studied viruses infecting different hosts such mammals, birds, fish and insects (Figure 4.9).

Although the methods that I have applied could not definitively identify that individual structural elements were or were not biologically functional, these methods consistently indicated that, for many of the analysed segments at least, a subset of the most conserved structures are very probably biologically functional. For instance, the finding that sites that are

predicted to be base paired show significant evidence of complementary coevolution in some of the datasets implies that mutations that disrupt base-pairing within some of the structures must lead to production of defective viruses that are purged from the population. A study by Gultayev et al 2016 made similar findings when it reported through an analysis of the HA segment that there existed subtype specific structural constraints within Influenza A viruses and that there was evidence of co-variation of base-paired nucleotides (Gultyaev *et al.*, 2016). This study also concluded that for evolution to have favoured this covariation these structural domains must play a role in viral fitness.

Through the analyses carried out here, inferences can be made on the impacts of secondary structures on the evolution of orthomyxoviruses: evolutionary imperatives to maintain functional secondary structures such as those which are involved in alternative splicing, can impose severe constraints on codon diversity, force dependence between the evolution of some base-paired sites, and result in variations in substitution rates between base-paired and non-base-paired sites. It is as yet unclear how these factors might affect the accuracy and reliability of sophisticated sequence analysis approaches that seek to extract detailed phylogeographical and epidemiological information from orthomyxovirus sequence datasets.

CHAPTER 5: CONCLUDING REMARKS

In recent years, global health challenges seem ever-changing and growing. There has been an upsurge in emerging and re-emerging viral pathogens coupled with increasing frequencies of antimicrobial resistance to existing treatment regimens (Marston *et al.*, 2014).

Properly understanding the evolutionary history and dissemination patterns of pathogens in the context of the environments in which they occur is crucial for understanding where, when and why pathogens emerge so that rational control strategies and new therapies can be developed and efficiently deployed at a global-scale (Gyawali and Taylor-Robinson, 2017).

I have demonstrated using phylodynamic approaches, that the transmission dynamics in space and time of an emerging epidemic, such as the African wave of the Influenza A/H1N1 2009 pandemic virus, can be deciphered with high precision from available geographical and temporal information relating to molecular A/H1N1 sequence data collected on the African continent that is comparable to approaches applied in other related studies(Merler *et al.*, 2011; de Silva, Ferguson and Fraser, 2012; He *et al.*, 2013; Watson *et al.*, 2015) . The application of such approaches could improve our understanding of the evolutionary dynamics of many other rapidly evolving single stranded RNA viruses that cause epidemics on the continent as has been demonstrated in other cases (Assiri *et al.*, 2013; Beck *et al.*, 2013; T. T.-Y. Lam *et al.*, 2013; du Plessis and Stadler, 2015; Rife *et al.*, 2017).

Multiple factors are usually involved in the spread of pathogens during pandemics (Talbi *et al.*, 2010; Lemey *et al.*, 2014; Nunes *et al.*, 2014; Gräf, Vrancken, Junqueira, *et al.*, 2015). My analysis tested several ecological, biological, geographical and soci-economic factors that may have impacted the dissemination pattern of the H1N1 pandemic virus within the African continent. Using the statistical models integrated within my Bayesian phylogenetics analyses, it was inferred that the greatest predictor of dissemination for the 2009 H1N1virus in Africa was geographical proximity and human mobility i.e. geographical distances between the sampling locations and air travel. This finding contributes to the emerging body of evidence suggesting fast human global movement and proximal locality to the foci of outbreaks of infectious dieases impacts infectious disease spread. My findings are consistent with previous reports that suggested air travel contributed to fast global dissemination of 2009 pandemic H1N1 virus (Tatem and Rogers, 2006; Neatherlin *et al.*, 2013; Lemey *et al.*, 2014). This may imply that control measures that focus on limiting travel to and from foci of the outbreaks may

be effective in mitigating the long distance spread of future eprodemics. Additionally, road networks provided accessibility and contributed to shorter distance regional spread of the 2014 ebola outbreak in West Africa (Valeri *et al.*, 2016; Jansen van Vuren *et al.*, 2019),. Although not directly comparable, my finding could inform additional strategies to those currently deployed as a matter of public-health policy to limit close interaction with infected victims that were implemented by West African governments during the 2016 Ebola virus outbreak by limiting flights to only essential ones to affected areas (Zinszer *et al.*, 2017).

Molecular sequences when used in combination with other meta data about the sequences may also yield useful insights into the biology of many viral pathogens (Mehle *et al.*, 2012; Beard *et al.*, 2014; Jones *et al.*, 2019). Through the application of computational methods, my studies have demonstrated that the diversification of orthomyxoviruses is influenced by processes including the rapid accumulation of point mutations, frequent reassortment and selection pressures acting on genome folding (the formation of biologically functional secondary structures at certain sites within the genomes of these viruses).

Reassortment analyses performed in this study show that Influenza virus diversity is greatly enhanced through reassortment. New strains that are generated through reassortment have the potential to generate regional epidemics or pandemics and, therefore, continuous monitoring of when and where reassortant lineages emerge should be a priority area of Influenza research and management (Furuse, Suzuki and Oshitani, 2010; Kiseleva *et al.*, 2012; Lycett *et al.*, 2012; Wille *et al.*, 2013). Timely information about possible new strains may inform the formulation of targeted prevention strategies such as either the diversion of additional resources to prevent the spread of these strains or the formulation of new vaccines that provide effective protection of populations at risk of infection (Argimón *et al.*, 2016; Hadfield *et al.*, 2018). Information on the geographical locations where the reassortment events occurred that generated the presently circulating reassortant viruses would be helpful in focussing surveillance efforts on geographical hotspots of reassortment. This would enhance timely containment of the spread of such strains before localized outbreaks have the opportunity to beome pandemics.

A clearer understanding of the biology of orthomyxoviruses in general should foster the design of new antiviral drugs or vaccines against these viruses . In this regard I explored the RNA secondary structure landscapes of vral genomes that are classified within the orthomyxovirus family. This family includes some important pathogens known to greatly impact the health of human and animal population. My findings suggest that there exist several biologically

functional conserved secondary structures within the genomes of orthomyxoviruses. The structured genome regions of these viruses could be further explored to yield useful leads for novel drug targets. Of particular note is the finding that the segment of orthomyxavirus genomes that is homologous to segment eight of Influenza A virus (which encodes non-structural genes), has several conserved structures across all five of the orthomyxovirus species that I examined.

The main limitation of nucleotide sequence analysis based studies such as I have carried out, is that the nucleotide sequences that are examined have usually been sampled by convenience. Therefore while the findings of studies such as mine may provide useful insights, in many cases these insights will not be globally-generalisable. For example, there are clear disparities in research and surveillance efforts in different regions of the world that have introduced substantial sampling biases into public genome sequence databases for almost all viral taxa: biases that in the case of my study may have impacted my assessment of both where in the world the different major influenza virus subtypes first arose, and where the major influenza virus reassortment events occurred.

Neverthless, the studies herein demonstrate that pathogen genomics-based analytical approaches are key to understanding the mechanisms that drive the evolution of rapidly evolving viral pathogens, and illuminate how these approaches could be leveraged to improve the management of these pathogens.

REFERENCES

Aamelfot, M. *et al.* (2015) 'Host tropism of infectious salmon anaemia virus in marine and freshwater fish species', *Journal of Fish Diseases*, 38(8), pp. 687–694. doi: 10.1111/jfd.12284.

Aamelfot, M., Dale, O. B. and Falk, K. (2014) 'Infectious salmon anaemia - pathogenesis and tropism', *Journal of Fish Diseases*, 37(4), pp. 291–307. doi: 10.1111/jfd.12225.

Abecasis, A. B., Pingarilho, M. and Vandamme, A. M. (2018) 'Phylogenetic analysis as a forensic tool in HIV transmission investigations', *Aids*, 32(5), pp. 543–554. doi: 10.1097/QAD.0000000000001728.

Aberer, A. J., Kobert, K. and Stamatakis, A. (2014) 'Exabayes: Massively parallel bayesian tree inference for the whole-genome era', *Molecular Biology and Evolution*, 31(10), pp. 2553–2556. doi: 10.1093/molbev/msu236.

Acevedo, A., Brodsky, L. and Andino, R. (2013) 'Mutational and fitness landscapes of an RNA virus revealed through population sequencing', *Nature*. Nature Publishing Group, 505(7485), pp. 686–690. doi: 10.1038/nature12861.

Amonsin, A. *et al.* (2010) 'Genetic characterization of 2008 reassortant influenza A virus (H5N1), Thailand.', *Virology journal*, 7, p. 233. doi: 10.1186/1743-422X-7-233.

Amore, G. *et al.* (2010) 'Multi-year evolutionary dynamics of West Nile virus in suburban Chicago, USA, 2005-2007', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1548). doi: 10.1098/rstb.2010.0054.

Ampofo, W. K. *et al.* (2015) 'Strengthening the influenza vaccine virus selection and development process: Report of the 3rd WHO Informal Consultation for Improving Influenza Vaccine Virus Selection held at WHO headquarters, Geneva, Switzerland, 1-3 April 2014.', *Vaccine*, 33(36), pp. 4368–82. doi: 10.1016/j.vaccine.2015.06.090.

Anderson-Lee, J. *et al.* (2016) 'Principles for Predicting RNA Secondary Structure Design Difficulty', *Journal of Molecular Biology*. The Authors, 428(5), pp. 748–757. doi: 10.1016/j.jmb.2015.11.013.

Antinori, E. B. De, Mccracken, J. P. and Widdowson, M. (2014) 'The Role of Temperature and Humidity on Seasonal Influenza in Tropical Areas : Guatemala , El Salvador and', 9(6), pp. 2008–2013. doi: 10.1371/journal.pone.0100659.

Archer, Brett N, Timothy, G. a, *et al.* (2012) 'Introduction of 2009 pandemic influenza A virus subtype H1N1 into South Africa: clinical presentation, epidemiology, and transmissibility of the first 100 cases.', *The Journal of infectious diseases*, 206 Suppl(Suppl 1), pp. S148-53. doi: 10.1093/infdis/jis583.

Archer, Brett N. *et al.* (2012) 'Introduction of 2009 pandemic influenza a virus subtype H1N1 into South Africa: Clinical presentation, epidemiology, and transmissibility of the first 100 cases', *Journal of Infectious Diseases*, 206(SUPPL.1). doi: 10.1093/infdis/jis583.

Archer, Brett N, Tempia, S., *et al.* (2012) 'Reproductive number and serial interval of the first wave of influenza A(H1N1)pdm09 virus in South Africa.', *PloS one*, 7(11), p. e49482. doi: 10.1371/journal.pone.0049482.

Argimón, S. *et al.* (2016) 'Microreact: visualizing and sharing data for genomic

epidemiology and phylogeography', *Microbial genomics*, 2(11), p. e000093. doi: 10.1099/mgen.0.000093.

Assiri, A. *et al.* (2013) 'Hospital outbreak of Middle East respiratory syndrome coronavirus.', *The New England journal of medicine*, 369(5), pp. 407–16. doi: 10.1056/NEJMoa1306742.

Austin, F. J. (1978) 'Johnston Atoll virus (Quaranfil group) from Ornithodoros capensis (Ixodoidea: Argasidae) infesting a Gannet colony in New Zealand', *American Journal of Tropical Medicine and Hygiene*, 27(5), pp. 1045–1048.

Ayres, D. L. *et al.* (2019) 'BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics', *Systematic Biology*, 68(6), pp. 1052–1061. doi: 10.1093/sysbio/syz020.

Azarian, T. *et al.* (2016) 'Genomic Epidemiology of Methicillin- Resistant Staphylococcus aureus in a Neonatal Intensive Care Unit', pp. 1–20. doi: 10.1371/journal.pone.0164397.

Baele, G. *et al.* (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29(9), pp. 2157–2167. doi: 10.1093/molbev/mss084.

Baele, G. *et al.* (2017) 'Emerging concepts of data integration in pathogen phylodynamics', *Systematic Biology*, 66(1), pp. e47–e65. doi: 10.1093/sysbio/syw054.

Baele, G. *et al.* (2018) 'Recent advances in computational phylodynamics', *Current Opinion in Virology*. Elsevier B.V., 31, pp. 24–32. doi: 10.1016/j.coviro.2018.08.009.

Bahl, J. *et al.* (2013) 'Influenza a virus migration and persistence in North American wild birds.', *PLoS pathogens*, 9(8), p. e1003570. doi: 10.1371/journal.ppat.1003570.

Baillie, G. J. *et al.* (2012) 'Evolutionary Dynamics of Local Pandemic H1N1 / 2009 Influenza Virus Lineages Revealed by Whole-Genome Analysis', *Journal of Virology*, 86(1), pp. 11–18. doi: 10.1128/JVI.05347-11.

Baker, S. F. *et al.* (2014) 'Influenza A and B Virus Intertypic Reassortment through Compatible Viral Packaging Signals', *Journal of Virology*, 88(18), pp. 10778–10791. doi: 10.1128/JVI.01440-14.

Balish, A. *et al.* (2009) 'NIH Public Access Author Manuscript Science. Author manuscript; available in PMC 2012 January 4. Published in final edited form as: Science . 2009 July 10; 325(5937): 197–201. doi:10.1126/science.1176225. Antigenic and Genetic Characteristics of the Early', *Science*, 325(5937), pp. 197–201. doi: 10.1126/science.1176225.Antigenic.

Bao, Y. *et al.* (2008) 'The influenza virus resource at the National Center for Biotechnology Information.', *Journal of virology*, 82(2), pp. 596–601. doi: 10.1128/JVI.02005-07.

Barakat, A. *et al.* (2012) '2009 Pandemic influenza A virus subtype H1N1 in Morocco, 2009-2010: epidemiology, transmissibility, and factors associated with fatal cases.', *The Journal of infectious diseases*, 206 Suppl(Suppl 1), pp. S94-100. doi: 10.1093/infdis/jis547.

Batts, W. N. *et al.* (2017) 'Molecular characterization of a novel orthomyxovirus from rainbow and steelhead trout (Oncorhynchus mykiss)', *Virus Research*. Elsevier B.V., 230, pp. 38–49. doi: 10.1016/j.virusres.2017.01.005.

Beard, R. *et al.* (2014) 'Generalized linear models for identifying predictors of the evolutionary diffusion of viruses.', *AMIA Joint Summits on Translational Science*

proceedings. AMIA Joint Summits on Translational Science, 2014, pp. 23–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25717395%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4333690.

Beck, A. *et al.* (2013) 'Phylogeographic Reconstruction of African Yellow Fever Virus Isolates Indicates Recent Simultaneous Dispersal into East and West Africa', *PLoS Neglected Tropical Diseases*, 7(3), p. e1910. doi: 10.1371/journal.pntd.0001910.

Bedford, T. *et al.* (2010) 'Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2)', *Plos Pathogens*, 6(5), p. doi: Artn E1000918\rDoi 10.1371/Journal.Ppat.1000918.

Bedford, T. *et al.* (2014) 'Integrating influenza antigenic dynamics with molecular evolution', *eLife*, 2014(3), pp. 1–26. doi: 10.7554/eLife.01914.

Bedford, T. *et al.* (2015) 'Global circulation patterns of seasonal influenza viruses vary with antigenic drift', *Nature*, 523(7559), pp. 217–20. doi: 10.1038/nature14460.

Belalov, I. S. and Lukashev, A. N. (2013) 'Causes and implications of codon usage bias in RNA viruses.', *PloS one*, 8(2), p. e56642. doi: 10.1371/journal.pone.0056642.

Bernhart, S. H. *et al.* (2008) 'RNAalifold: improved consensus structure prediction for RNA alignments.', *BMC bioinformatics*, 9, p. 474. doi: 10.1186/1471-2105-9-474.

Bhatt, S. *et al.* (2013) 'The evolutionary dynamics of influenza A virus adaptation to mammalian hosts The evolutionary dynamics of influenza A virus adaptation to mammalian hosts'.

Biek, R. *et al.* (2015) 'Measurably evolving pathogens in the genomic era', *Trends in Ecology and Evolution*, 30(6). doi: 10.1016/j.tree.2015.03.009.

Bielejec, F. *et al.* (2011) 'SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics', *Bioinformatics*, 27(20), pp. 2910–2912. doi: 10.1093/bioinformatics/btr481.

Biggerstaff, M. *et al.* (2014) 'Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature.', *BMC infectious diseases*, 14(1), p. 480. doi: 10.1186/1471-2334-14-480.

Bindewald, E. and Shapiro, B. (2006) 'RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers', *Rna*, pp. 342–352. doi: 10.1261/rna.2164906.uses.

Bloom, D. E., Black, S. and Rappuoli, R. (2017) 'Emerging infectious diseases: A proactive approach', *Proceedings of the National Academy of Sciences*, 114(16), pp. 4055–4059. doi: 10.1073/pnas.1701410114.

Bloomquist, E. W., Lemey, P. and Suchard, M. A. (2010) 'Three roads diverged? Routes to phylogeographic inference', *Trends in Ecology & Evolution*. Elsevier Ltd, 25(11), pp. 626–632. doi: 10.1016/j.tree.2010.08.010.

Boni, M. F. *et al.* (2010) 'Guidelines for identifying homologous recombination events in influenza A virus', *PLoS ONE*, 5(5), pp. 1–11. doi: 10.1371/journal.pone.0010434.

Bromham, L. *et al.* (2018) 'Bayesian molecular dating: opening up the black box', *Biological Reviews*, 93(2), pp. 1165–1191. doi: 10.1111/brv.12390.

Brunker, K. *et al.* (2012) 'Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model', *Parasitology*, pp. 1–15. doi: 10.1017/S003118201200090X.

Brunner, F. S. *et al.* (2019) 'The diversity of eco-evolutionary dynamics: Comparing the feedbacks between ecology and evolution across scales', *Functional Ecology*, 33(1), pp. 7–12. doi: 10.1111/1365-2435.13268.

Bulimo, W. D. *et al.* (2012) 'Molecular characterization and phylogenetic analysis of the hemagglutinin 1 protein of human influenza A virus subtype H1N1 circulating in Kenya during 2007-2008.', *The Journal of infectious diseases*, 206 Suppl(Suppl 1), pp. S46-52. doi: 10.1093/infdis/jis586.

Buratti, E. and Baralle, F. E. (2004) 'Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process MINIREVIEW Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process', *Molecular and Cellular Biology*, 24(24), pp. 10505–10514. doi: 10.1128/MCB.24.24.10505.

Burkhardt, C. *et al.* (2014) 'Structural constraints in the packaging of bluetongue virus genomic segments', *Journal of General Virology*, 95, pp. 2240–2250. doi: 10.1099/vir.0.066647-0.

Byrd-Leotis, L. *et al.* (2015) 'Influenza Hemagglutinin (HA) Stem Region Mutations That Stabilize or Destabilize the Structure of Multiple HA Subtypes.', *Journal of virology*, 89(8), pp. 4504–16. doi: 10.1128/JVI.00057-15.

Campbell, F. *et al.* (2018) 'When are pathogen genome sequences informative of transmission events?', *Plos One*, pp. 1–21. doi: 10.1371/journal.ppat.1006885.

Carter, R. and Mendis, K. N. (2002) 'Evolutionary and Historical Aspects of the Burden of Malaria Evolutionary and Historical Aspects of the Burden of Malaria', *Clin. Microbiol. Rev*, 15(4), pp. 564–594. doi: 10.1128/CMR.15.4.564.

Carvalho, L. M. *et al.* (2015) 'Spatio-temporal Dynamics of Foot-and-Mouth Disease Virus in South America', pp. 1–21. Available at: http://arxiv.org/abs/1505.01105.

Castro-Nallar, E. *et al.* (2011) 'Molecular phylodynamics and protein modeling of infectious salmon anemia virus (ISAV)', *BMC Evolutionary Biology*, 11(1). doi: 10.1186/1471-2148-11-349.

Castro-Nallar, E. *et al.* (2012) 'The evolution of HIV: Inferences using phylogenetics', *Molecular Phylogenetics and Evolution*. Elsevier Inc., 62(2), pp. 777–792. doi: 10.1016/j.ympev.2011.11.019.

Chen, G. W. and Shih, S. R. (2009) 'Genomic signatures of influenza A pandemic (H1N1) 2009 virus', *Emerging Infectious Diseases*, 15(12), pp. 1897–1903. doi: 10.3201/eid1512.090845.

Chen, L. *et al.* (2016) 'Is a highly pathogenic avian influenza virus H5N1 fragment recombined in PB1 the key for the epidemic of the novel AIV H7N9 in China, 2013?', *International Journal of Infectious Diseases*, 43, pp. 85–89. doi: 10.1016/j.ijid.2016.01.002.

Chen, R. and Holmes, E. C. (2008) 'The evolutionary dynamics of human influenza B virus.', *Journal of molecular evolution*, 66(6), pp. 655–63. doi: 10.1007/s00239-008-9119-z.

Chevenet, F. *et al.* (2013) 'Searching for virus phylotypes', *Bioinformatics*, 29(5), pp. 561–570. doi: 10.1093/bioinformatics/btt010.

Christman, M. C. *et al.* (2011) 'Pandemic (H1N1) 2009 virus revisited: an evolutionary retrospective.', *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 11(5), pp. 803–11. doi: 10.1016/j.meegid.2011.02.021.

Cloete, L. J. *et al.* (2014) 'The influence of secondary structure , selection and recombination on r ubella virus nucleotide substitution rate estimates', pp. 1–12.

Compans, R. and Oldstone, M. (2014) *Influenza Pathogenesis and Control - Current Topics in Microbiology and Immunology*.

Conroy, G. C. *et al.* (2008) 'Google Earth, GIS, and the Great Divide: a new and simple method for sharing paleontological data.', *Journal of human evolution*, 55(4), pp. 751–5. doi: 10.1016/j.jhevol.2008.03.001.

Contreras-Gutiérrez, M. A. *et al.* (2017) 'Corrigendum to "Sinu virus, a novel and divergent orthomyxovirus related to members of the genus Thogotovirus isolated from mosquitoes in Colombia" [Virology 501 (2017) 166–175] (S0042682216303683)(10.1016/j.virol.2016.11.014)', *Virology*. Elsevier, 503(November 2016), p. 114. doi: 10.1016/j.virol.2017.02.005.

Da Costa, B. *et al.* (2015) 'Temperature-Sensitive Mutants in the Influenza A Virus RNA Polymerase: Alterations in the PA Linker Reduce Nuclear Targeting of the PB1-PA Dimer and Result in Viral Attenuation.', *Journal of virology*. American Society for Microbiology, 89(12), pp. 6376–90. doi: 10.1128/JVI.00589-15.

Cottam, E. M. *et al.* (2008) 'Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus.', *Proceedings. Biological sciences / The Royal Society*, 275(1637), pp. 887–895. doi: 10.1098/rspb.2007.1442.

Cottet, L. *et al.* (2011) 'Infectious salmon anemia virus—Genetics and pathogenesis', *Virus Research*. Elsevier B.V., 155(1), pp. 10–19. doi: 10.1016/j.virusres.2010.10.021.

Cros, J. F. and Palese, P. (2003) 'Trafficking of viral genomic RNA into and out of the nucleus: influenza, Thogoto and Borna disease viruses', *Virus Research*, 95(1–2), pp. 3–12. doi: 10.1016/S0168-1702(03)00159-X.

Dela-Moss, L. I., Moss, W. N. and Turner, D. H. (2014a) 'Identification of conserved RNA secondary structures at influenza B and C splice sites reveals similarities and differences between influenza A, B, and C.', *BMC research notes*, 7, p. 22. doi: 10.1186/1756-0500-7-22.

Dela-Moss, L. I., Moss, W. N. and Turner, D. H. (2014b) 'Identification of conserved RNA secondary structures at influenza B and C splice sites reveals similarities and differences between influenza A, B, and C.', *BMC research notes*, 7, p. 22. doi: 10.1186/1756-0500-7-22.

Desselberger, U. *et al.* (1978) 'Biochemical evidence that "new" influenza virus strains in nature may arise by recombination (reassortment).', *Proceedings of the National Academy of Sciences of the United States of America*, 75(7), pp. 3341–3345. doi: 10.1073/pnas.75.7.3341.

Devaux, C. A. (2012) 'Emerging and re-emerging viruses: A global challenge illustrated by

Chikungunya virus outbreaks', *World Journal of Virology*, 1(1), p. 11. doi: 10.5501/wjv.v1.i1.11.

Dia, N. *et al.* (2013) 'A Subregional Analysis of Epidemiologic and Genetic Characteristics of Influenza A(H1N1)pdm09 in Africa: Senegal, Cape Verde, Mauritania, and Guinea, 2009-2010', *American Journal of Tropical Medicine and Hygiene*, 88(5), pp. 946–953. doi: 10.4269/ajtmh.12-0401.

Drummond, A. J. *et al.* (2005) 'Bayesian coalescent inference of past population dynamics from molecular sequences', *Molecular Biology and Evolution*, 22(5), pp. 1185–1192. doi: 10.1093/molbev/msi103.

Drummond, A. J. *et al.* (2006) 'Relaxed phylogenetics and dating with confidence', *PLoS Biol.* , 4, p. e88. Available at: http://dx.doi.org/10.1371/journal.pbio.0040088.

Drummond, A. J. *et al.* (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29(8), pp. 1969–1973. doi: 10.1093/molbev/mss075.

Drummond, A. J. and Bouckaert, R. R. (2014) 'Bayesian evolutionary analysis with BEAST 2', p. 249.

Drummond, A. J. and Rambaut, A. (2007) 'BEAST: Bayesian evolutionary analysis by sampling trees', *BMC Evol. Biol.* , 7, p. 214. Available at: http://dx.doi.org/10.1186/1471-2148-7-214.

Drummond, A. J., Rambaut, A. and Xie, W. (2011) 'Bayesian Skyline Plot', *Beast*.

Duchêne, S., Holmes, E. C. and Ho, S. Y. W. (2014) 'Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates.', *Proceedings. Biological sciences / The Royal Society*, 281(1786), pp. 20140732-. doi: 10.1098/rspb.2014.0732.

Dudas, G. *et al.* (2015) 'Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1-PB2-HA Gene Complex', *Molecular Biology and Evolution*, 32(1), pp. 162–172. doi: 10.1093/molbev/msu287.

Dudas, G. *et al.* (2017) 'Virus genomes reveal factors that spread and sustained the Ebola epidemic', *Nature*, 544(7650), pp. 309–315. doi: 10.1038/nature22040.

Duke-Sylvester, S. M., Biek, R. and Real, L. a (2013) 'Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614), p. 20120194. doi: 10.1098/rstb.2012.0194.

Edgar, Robert C (2004) 'MUSCLE: a multiple sequence alignment method with reduced time and space complexity.', *BMC bioinformatics*, 5, p. 113. doi: 10.1186/1471-2105-5-113.

Edgar, R C (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Res.* , 32, pp. 1792–1797. Available at: http://dx.doi.org/10.1093/nar/gkh340.

Edgar, R. C. (2010) 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, 26(19), pp. 2460–2461. doi: 10.1093/bioinformatics/btq461.

Elbe, S. and Buckland-Merrett, G. (2017) 'Data, disease and diplomacy: GISAID's innovative contribution to global health', *Global Challenges*, 1(1), pp. 33–46. doi: 10.1002/gch2.1018.

Engel, G. a *et al.* (2013) 'Zoonotic simian foamy virus in Bangladesh reflects diverse patterns of transmission and co-infection', *Emerging Microbes & Infections*, 2(9), p. e58. doi: 10.1038/emi.2013.60.

Falk, K. *et al.* (1997) 'Characterization of infectious salmon anemia virus, an orthomyxo-like virus isolated from Atlantic salmon (Salmo salar L.).', *Journal of virology*, 71(12), pp. 9016–9023.

Faria, N. R. *et al.* (2011) 'Toward a quantitative understanding of viral phylogeography', *Current Opinion in Virology*, 1(5). doi: 10.1016/j.coviro.2011.10.003.

Faria, N. R. *et al.* (2013) 'Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614), p. 20120196. doi: 10.1098/rstb.2012.0196.

Faria, N. R. *et al.* (2014) 'HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations', *Science*, 346(6205), pp. 56–61. doi: 10.1126/science.1256739.

Faria, N. R. *et al.* (2016) 'Zika virus in the Americas: Early epidemiological and genetic findings', *Science*, 5036(March), pp. 1–9. doi: 10.1126/science.aaf5036.

Fields, B. N., Knipe, D. M. and Howley, P. M. (2007) *Fields Virology, 5th Edition*, *Fields Virology*. Available at: http://www.loc.gov/catdir/toc/ecip072/2006032230.html.

Fischer, W. A. *et al.* (2014) 'Global Burden of Influenza as a Cause of Cardiopulmonary Morbidity and Mortality.', *Global heart*. World Heart Federation (Geneva), 9(3), pp. 325–336. doi: 10.1016/j.gheart.2014.08.004.

Flouri, T. *et al.* (2015) 'The phylogenetic likelihood library', *Systematic Biology*, 64(2), pp. 356–362. doi: 10.1093/sysbio/syu084.

Foni, E. *et al.* (2017) 'Influenza D in Italy: towards a better understanding of an emerging viral infection in swine', *Scientific Reports*. Springer US, 7(1), p. 11660. doi: 10.1038/s41598-017-12012-3.

Forrest, H. L. and Webster, R. G. (2010) 'Perspectives on influenza evolution and the role of research.', *Animal Health Research Reviews*, 11(1), pp. 3–18. doi: 10.1017/S1466252310000071.

Fourment, M. and Darling, A. E. (2018) 'Local and relaxed clocks: The best of both worlds', *PeerJ*, 2018(7). doi: 10.7717/peerj.5140.

Freire, C. C. M., Iamarino, A., Soumaré, P. O. L., Faye, O., Sall, A. a., Guan, Y., *et al.* (2015) 'Guidelines for identifying homologous recombination events in influenza A virus', *PLoS ONE*. Genome Biology, 5(5), pp. 1–11. doi: 10.1371/journal.pone.0010434.

Freire, C. C. M., Iamarino, A., Soumaré, P. O. L., Faye, O., Sall, A. a. and Zanotto, P. M. a. (2015) 'Reassortment and distinct evolutionary dynamics of Rift Valley Fever virus genomic segments', *Scientific Reports*. Nature Publishing Group, 5(May), p. 11353. doi: 10.1038/srep11353.

Freyhult, E., Moulton, V. and Gardner, P. (2005) 'Predicting RNA structure using mutual information', *Applied Bioinformatics*, 4(1), pp. 53–59. doi: 10.2165/00822942-200504010-00006.

Frost, S. D. W. *et al.* (2015) 'Eight challenges in phylodynamic inference', *Epidemics*. Elsevier B.V., 10, pp. 88–92. doi: 10.1016/j.epidem.2014.09.001.

Fuller, T. L. *et al.* (2013) 'Predicting hotspots for influenza virus reassortment', *Emerging Infectious Diseases*, 19(4), pp. 581–588. doi: 10.3201/eid1904.120903.

Furuse, Y., Suzuki, A. and Oshitani, H. (2010) 'Reassortment between swine influenza A viruses increased their adaptation to humans in pandemic H1N1/09', *Infection, Genetics and Evolution*. Elsevier B.V., 10(4), pp. 569–574. doi: 10.1016/j.meegid.2010.01.010.

Gagné, N. and LeBlanc, F. (2017) 'Overview of infectious salmon anaemia virus (ISAV) in Atlantic Canada and first report of an ISAV North American-HPR0 subtype', *Journal of Fish Diseases*, (May). doi: 10.1111/jfd.12670.

Galiano, M. *et al.* (2011) 'Evolutionary pathways of the pandemic influenza A (H1N1) 2009 in the UK.', *PLoS ONE*, 6(8), p. e23779. doi: 10.1371/journal.pone.0023779.

Gao, Q. and Palese, P. (2009) 'Rewiring the RNAs of influenza virus to prevent reassortment', *Proceedings of the National Academy of Sciences of the United States of America*, 106(37), pp. 15891–15896. doi: 10.1073/pnas.0908897106.

García, K., Díaz, A. and Navarrete, A. (2013) 'New strategies for control, prevention and treatment of ISA virus in aquaculture', *... and Strategies for ...*, 1, pp. 587–597. Available at: http://www.formatex.info/microbiology4/vol1/587-597.pdf.

Gardner, L. and Sarkar, S. (2013) 'A global airport-based risk model for the spread of dengue infection via the air transport network.', *PloS one*, 8(8), p. e72129. doi: 10.1371/journal.pone.0072129.

Garten, R. J. *et al.* (2009) 'Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans.', *Science (New York, N.Y.)*, 325(5937), pp. 197–201. doi: 10.1126/science.1176225.

Gebreyes, W. A. *et al.* (2014) 'The global one health paradigm: challenges and opportunities for tackling infectious diseases at the human, animal, and environment interface in low-resource settings.', *PLoS neglected tropical diseases*, 8(11), p. e3257. doi: 10.1371/journal.pntd.0003257.

Gerber, M. *et al.* (2014) 'Selective packaging of the influenza A genome and consequences for genetic reassortment', *Trends in Microbiology*. Elsevier Ltd, 22(8), pp. 446–455. doi: 10.1016/j.tim.2014.04.001.

Gerhardt, G. J. L. *et al.* (2013) 'Triplet entropy analysis of hemagglutinin and neuraminidase sequences measures influenza virus phylodynamics.', *Gene*, 528(2), pp. 277–81. doi: 10.1016/j.gene.2013.06.060.

Gessner, B. D., Shindo, N. and Briand, S. (2011) 'Seasonal influenza epidemiology in sub-Saharan Africa: a systematic review.', *The Lancet. Infectious diseases*, 11(3), pp. 223–35. doi: 10.1016/S1473-3099(11)70008-1.

Giardina, F. *et al.* (2017) 'Inference of Transmission Network Structure from HIV Phylogenetic Trees', *PLoS Computational Biology*, 13(1), pp. 1–22. doi: 10.1371/journal.pcbi.1005316.

Gibbs, A. J., Armstrong, J. S. and Downie, J. C. (2009) 'From where did the 2009 "swine-

origin" influenza A virus (H1N1) emerge?', *Virology Journal*, 6. doi: 10.1186/1743-422X-6-207.

Gill, M. S. *et al.* (2017) 'A relaxed directional random walk model for phylogenetic trait evolution', *Systematic Biology*, 66(3), pp. 299–319. doi: 10.1093/sysbio/syw093.

Gilsdorf, A., Morgan, D. and Leitmeyer, K. (2012) 'Guidance for contact tracing of cases of Lassa fever, Ebola or Marburg haemorrhagic fever on an airplane: results of a European expert consultation.', *BMC public health*, 12(1), p. 1014. doi: 10.1186/1471-2458-12-1014.

Gire, S. K. *et al.* (2014) 'Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak', *Science*, 345(6202), pp. 1369–72. doi: 10.1126/science.1259657.

Godoy, M. G. *et al.* (2008) 'First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed Atlantic salmon (Salmo salar) in Chile', *BMC Veterinary Research*, 4(1), p. 28. doi: 10.1186/1746-6148-4-28.

Gog, J. R. *et al.* (2014) 'Spatial Transmission of 2009 Pandemic Influenza in the US.', *PLoS computational biology*, 10(6), p. e1003635. doi: 10.1371/journal.pcbi.1003635.

Goodacre, S. W. (2013) 'Cost utility of rapid polymerase chain reaction-based influenza testing: What should a clinician make of this analysis?', *Annals of Emergency Medicine*. Elsevier Inc., 62(1), pp. 89–90. doi: 10.1016/j.annemergmed.2013.01.027.

Gouy, M., Guindon, S. and Gascuel, O. (2010) 'SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building', *Molecular Biology and Evolution*, 27(2), pp. 221–224. doi: 10.1093/molbev/msp259.

Grabenstein, J. D. and Nevin, R. L. (2006) 'Mass immunization programs: principles and standards.', *Current topics in microbiology and immunology*, 304, pp. 31–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16989263.

Grad, Y. H. and Lipsitch, M. (2014) 'Epidemiologic data and pathogen genome sequences: a powerful synergy for public health', *Genome Biology*, 15(11), p. 538. doi: 10.1186/s13059-014-0538-4.

Gräf, T., Vrancken, B., Maletich Junqueira, D., *et al.* (2015) 'Contribution of Epidemiological Predictors in Unraveling the Phylogeographic History of HIV-1 Subtype C in Brazil', *Journal of Virology*, 89(24), pp. 12341–12348. doi: 10.1128/JVI.01681-15.

Gräf, T., Vrancken, B., Junqueira, D. M., *et al.* (2015) 'The contribution of epidemiological predictors in unravelling the phylogeographic history of HIV-1 subtype C in Brazil', *Journal of Virology*, (September), p. JVI.01681-15. doi: 10.1128/JVI.01681-15.

Grassly, N. C. and Fraser, C. (2006) 'Seasonal infectious disease epidemiology', *Proceedings of the Royal Society B: Biological Sciences*, 273(April), pp. 2541–2550. doi: 10.1098/rspb.2006.3604.

Gray, R. R. and Salemi, M. (2012) 'Integrative molecular phylogeography in the context of infectious diseases on the human-animal interface.', *Parasitology*, 139(14), pp. 1939–51. doi: 10.1017/S0031182012001102.

Grear, D. A. *et al.* (2018) 'Inferring epidemiologic dynamics from viral evolution: 2014–2015 Eurasian/North American highly pathogenic avian influenza viruses exceed transmission threshold, R 0 = 1, in wild birds and poultry in North America', *Evolutionary*

Applications, 11(4), pp. 547–557. doi: 10.1111/eva.12576.

Greenbaum, B. D. *et al.* (2012) 'Viral reassortment as an information exchange between viral segments', *Proceedings of the National Academy of Sciences*, 109(9), pp. 3341–3346. doi: 10.1073/pnas.1113300109.

Grenfell, B. T. (2004) 'Unifying the epidemiological and evolutionary dynamics of pathogens', *Science*. , 303, pp. 327–332. Available at: http://dx.doi.org/10.1126/science.1090727.

Grenfell, B. T. *et al.* (2004) 'Unifying the Epidemiological and Evolutionary Dynamics of Pathogens', *Science*, 303(5656). doi: 10.1126/science.1090727.

Gultyaev, Alexander P *et al.* (2014) 'RNA structural constraints in the evolution of the influenza A virus genome NP segment.', *RNA biology*. Landes Bioscience, 11(7), pp. 942–52. doi: 10.4161/rna.29730.

Gultyaev, A P *et al.* (2014) 'RNA structural constraints in the evolution of the influenza A virus genome NP segment', *RNA Biol*, 11(7), pp. 942–952. doi: 10.4161/rna.29730.

Gultyaev, A. P. *et al.* (2016) 'Subtype-specific structural constraints in the evolution of influenza A virus hemagglutinin genes', *Scientific Reports*, 6(November), p. 38892. doi: 10.1038/srep38892.

Gultyaev, A. P., Fouchier, R. A. M. and Olsthoorn, R. C. L. (2010) 'Influenza virus RNA structure: unique and common features.', *International reviews of immunology*, 29(6), pp. 533–56. doi: 10.3109/08830185.2010.507828.

Gultyaev, A. P. and Olsthoorn, R. C. (2010) 'A family of non-classical pseudoknots in influenza A and B viruses', *RNA Biol*, 7(2), pp. 125–129. doi: 10.4161/rna.7.2.11287.

Guu, T. S. Y. *et al.* (2008) 'Mapping the domain structure of the influenza A virus polymerase acidic protein (PA) and its interaction with the basic protein 1 (PB1) subunit', *Virology*, 379(1), pp. 135–142. doi: 10.1016/j.virol.2008.06.022.

Gyawali, N. and Taylor-Robinson, A. (2017) 'Confronting the Emerging Threat to Public Health in Northern Australia of Neglected Indigenous Arboviruses', *Tropical Medicine and Infectious Disease*, 2(4), p. 55. doi: 10.3390/tropicalmed2040055.

Hadfield, J. *et al.* (2018) 'NextStrain: Real-time tracking of pathogen evolution', *Bioinformatics*, 34(23), pp. 4121–4123. doi: 10.1093/bioinformatics/bty407.

Hannoun, C. (2013) 'The evolving history of influenza viruses and influenza vaccines.', *Expert Review of Vaccines*, 12(9), pp. 1085–94. doi: 10.1586/14760584.2013.824709.

Hause, B. *et al.* (2014) 'Characterization of a novel influenza virus strain in cattle and swine: proposal for a new genus in the Orthomyxoviridae family', *MBio*, 5(2), pp. 1–10. doi: 10.1128/mBio.00031-14.

Hay, S. I. *et al.* (2013) 'Global mapping of infectious disease Global mapping of infectious disease', (February).

He, D. *et al.* (2013) 'Patterns of spread of influenza A in Canada.', *Proceedings. Biological sciences / The Royal Society*, 280(1770), p. 20131174. doi: 10.1098/rspb.2013.1174.

He, L. and Zhu, J. (2015) 'Computational tools for epitope vaccine design and evaluation',

Current Opinion in Virology. Elsevier B.V., 11, pp. 103–112. doi: 10.1016/j.coviro.2015.03.013.

Hedge, J., Rambaut, A. and Lycett, S. J. (2013) 'Data from: Real-time characterization of the molecular epidemiology of an influenza pandemic', *Biology Letters*, 2010(April 2010). doi: doi:10.5061/dryad.jm858.

Heller, R., Chikhi, L. and Siegismund, H. R. (2013) 'The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History', *PLoS ONE*, 8(5). doi: 10.1371/journal.pone.0062992.

van Hemert, F., van der Kuyl, A. C. and Berkhout, B. (2016) 'Impact of the biased nucleotide composition of viral RNA genomes on RNA structure and codon usage', *Journal of General Virology*, 97(10), pp. 2608–2619. doi: 10.1099/jgv.0.000579.

Hemphill, M. L. *et al.* (1993) 'Antigenic and genetic analyses of influenza type B viruses isolated in Russia, 1987-91', *Epidemiol Infect*, 111(3), pp. 539–546. doi: 10.1017/S0950268800057265.

Heraud, J.-M. *et al.* (2012) 'Spatiotemporal circulation of influenza viruses in 5 African countries during 2008-2009: a collaborative study of the Institut Pasteur International Network.', *The Journal of infectious diseases*, 206 Suppl(Suppl 1), pp. S5-13. doi: 10.1093/infdis/jis541.

Hilleman, M. R. (2002) 'Realities and enigmas of human viral influenza: Pathogenesis, epidemiology and control', *Vaccine*, 20(25–26), pp. 3068–3087. doi: 10.1016/S0264-410X(02)00254-2.

Ho, S. Y. W. *et al.* (2011) 'Time-dependent rates of molecular evolution', *Mol Ecol*, 20(15), pp. 3087–3101. doi: DOI 10.1111/j.1365-294X.2011.05178.x.

HO, S. Y. W. and SHAPIRO, B. (2011) 'Skyline-plot methods for estimating demographic history from nucleotide sequences', *Molecular Ecology Resources*, 11(3), pp. 423–434. doi: 10.1111/j.1755-0998.2011.02988.x.

Hofacker, I. L., Fekete, M. and Stadler, P. F. (2002) 'Secondary structure prediction for aligned RNA sequences', *Journal of Molecular Biology*, 319(5), pp. 1059–1066. doi: 10.1016/S0022-2836(02)00308-X.

Hoft, D. F. and Belshe, R. B. (2004) 'The Genetic Archaeology of Influenza', *New England Journal of Medicine*, 351(24), pp. 2550–2551. doi: 10.1056/NEJMcibr043708.

Holmes, E. C. *et al.* (2005) 'Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses', *PLoS Biology*, 3(9), pp. 1579–1589. doi: 10.1371/journal.pbio.0030300.

Holmes, E. C. and Grenfell, B. T. (2009) 'Discovering the phylodynamics of RNA viruses', *PLoS Computational Biology*, 5(10). doi: 10.1371/journal.pcbi.1000505.

Hughes, G. J. *et al.* (2009) 'Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom', *PLoS Pathogens*, 5(9). doi: 10.1371/journal.ppat.1000590.

Hyphy, T. (no date) 'Complementary coevolution between paired nucleotides'.

Iglesias, N. G. and Gamarnik, A. V. (2014) 'Dynamic RNA structures in the dengue virus genome', *RNA Biology*, 8(2), pp. 249–257. doi: 10.4161/rna.8.2.14992.

Jaeger, J. A., Turner, D. H. and Zuker, M. (1989) 'Improved predictions of secondary structures for RNA', *Biochemistry*, 86, pp. 7706–7710. doi: 10.1073/pnas.86.20.7706.

Jansen van Vuren, P. *et al.* (2019) 'Phylodynamic Analysis of Ebola Virus Disease Transmission in Sierra Leone', *Viruses*, 11(1), p. 71. doi: 10.3390/v11010071.

Jardetzky, T. S. and Lamb, R. A. (2004) 'Virology: a class act.', *Nature*, 427(6972), pp. 307–8. doi: 10.1038/427307a.

Jombart, T. *et al.* (2009) 'Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic.', pp. 1–15. doi: 10.1371/currents.RRN1026.Authors.

Jones, S. *et al.* (2019) 'Evolutionary, genetic, structural characterization and its functional implications for the influenza A (H1N1) infection outbreak in India from 2009 to 2017', *Scientific reports*. Springer US, 9(1), p. 14690. doi: 10.1038/s41598-019-51097-w.

Joseph, U. *et al.* (2015) 'Adaptation of Pandemic H2N2 Influenza A Viruses in Humans', *Journal of Virology*, 89(4), pp. 2442–2447. doi: 10.1128/JVI.02590-14.

Joseph, U. *et al.* (2017) 'The ecology and adaptive evolution of influenza A interspecies transmission', *Influenza and other Respiratory Viruses*, 11(1), pp. 74–84. doi: 10.1111/irv.12412.

Kao, R. R. *et al.* (2014) 'Supersize me: How whole-genome sequencing and big data are transforming epidemiology', *Trends in Microbiology*. Elsevier Ltd, 22(5), pp. 282–291. doi: 10.1016/j.tim.2014.02.011.

Karcher, M. D. *et al.* (2016) 'Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference', *PLoS Computational Biology*, 12(3), pp. 1–19. doi: 10.1371/journal.pcbi.1004789.

Katz, M. a. *et al.* (2012) 'Influenza in Africa: Uncovering the epidemiology of a long-overlooked disease', *Journal of Infectious Diseases*, 206(SUPPL.1), pp. 4–7. doi: 10.1093/infdis/jis548.

Kellam, P. *et al.* (2003) 'Viral bioinformatics: computational views of host and pathogen.', *Novartis Foundation symposium*, 254, pp. 234–247; discussion 247-252.

Kibenge, F. S. *et al.* (2009) 'Infectious salmon anaemia virus (ISAV) isolated from the ISA disease outbreaks in Chile diverged from ISAV isolates from Norway around 1996 and was disseminated around 2005, based on surface glycoprotein gene sequences', *Virology Journal*, 6(June), pp. 1–16. doi: 10.1186/1743-422X-6-88.

Kibenge, F. S. B. *et al.* (2004) 'Infectious salmon anemia virus: causative agent, pathogenesis and immunity', *Animal Health Research Reviews*, 5(01), pp. 65–78. doi: 10.1079/AHRR200461.

Kibenge, M. J. *et al.* (2016) 'Discovery of variant infectious salmon anaemia virus (ISAV) of European genotype in British Columbia, Canada', *Virology Journal*. Virology Journal, 13(1), p. 3. doi: 10.1186/s12985-015-0459-1.

Kierzek, E. *et al.* (2006) 'Nearest neighbor parameters for Watson-Crick complementary heteroduplexes formed between 2???-O-methyl RNA and RNA oligonucleotides', *Nucleic Acids Research*, 34(13), pp. 3609–3614. doi: 10.1093/nar/gkl232.

King, A. M. Q. *et al.* (2018) 'Changes to taxonomy and the International Code of Virus

Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018)', *Archives of Virology*. Springer Vienna, 163(9), pp. 2601–2631. doi: 10.1007/s00705-018-3847-1.

Kiseleva, I. *et al.* (2012) 'Possible outcomes of reassortment in vivo between wild type and live attenuated influenza vaccine strains', *Vaccine*. Elsevier Ltd, 30(51), pp. 7395–7399. doi: 10.1016/j.vaccine.2012.09.076.

Klepac, P. *et al.* (2014) 'Six challenges in the eradication of infectious diseases', *Epidemics*. Elsevier B.V., 10, pp. 97–101. doi: 10.1016/j.epidem.2014.12.001.

Kobasa, D. and Kawaoka, Y. (2005) 'Emerging influenza viruses: past and present', *Curr Mol Med*, 5(8), pp. 791–803. doi: 10.2174/156652405774962281.

Kobayashi, Y., Dadonaite, B., Doremalen, N. Van, *et al.* (2016) 'Computational and molecular analysis of conserved influenza A virus RNA secondary structures involved in infectious virion production', *RNA biology*. Taylor & Francis, 13(9), pp. 883–894. doi: 10.1080/15476286.2016.1208331.

Kobayashi, Y., Dadonaite, B., van Doremalen, N., *et al.* (2016) 'Computational and molecular analysis of conserved influenza A virus RNA secondary structures involved in infectious virion production', *RNA Biology*. Taylor & Francis, 13(9), pp. 883–894. doi: 10.1080/15476286.2016.1208331.

Koelle, K. *et al.* (2011) 'A dimensionless number for understanding the evolutionary dynamics of antigenically variable RNA viruses', *Proceedings of the Royal Society B: Biological Sciences*, 278(1725). doi: 10.1098/rspb.2011.0435.

Kosakovsky Pond, S. L. *et al.* (2006) 'GARD: A genetic algorithm for recombination detection', *Bioinformatics*, 22(24), pp. 3096–3098. doi: 10.1093/bioinformatics/btl474.

Kosakovsky Pond, S. L., Frost, S. D. W. and Muse, S. V. (2005) 'HyPhy: Hypothesis testing using phylogenies', *Bioinformatics*, 21(5), pp. 676–679. doi: 10.1093/bioinformatics/bti079.

Kosoy, O. I. *et al.* (2015) 'Novel Thogotovirus associated with febrile illness and death, united states, 2014', *Emerging Infectious Diseases*, 21(5), pp. 760–764. doi: 10.3201/eid2105.150150.

Kühnert, D. *et al.* (2014) 'Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model', *Journal of the Royal Society Interface*, 11(94). doi: 10.1098/rsif.2013.1106.

Kühnert, D., Wu, C.-H. and Drummond, A. J. (2011) 'Phylogenetic and epidemic modeling of rapidly evolving infectious diseases', *Infection, Genetics and Evolution*, 11(8), pp. 1825–1841. doi: 10.1016/j.meegid.2011.08.005.

Kuno, G. *et al.* (2001) 'Phylogeny of Thogoto virus', *Virus Genes*, 23(2), pp. 211–214. doi: 10.1023/A:1011873028144.

Lam, T. T.-Y. *et al.* (2011) 'Reassortment events among swine influenza A viruses in China: implications for the origin of the 2009 influenza pandemic', *Journal of Virology*, 85(19), pp. 10279–10285. doi: 10.1128/JVI.05262-11.

Lam, T. T.-Y. *et al.* (2012) 'Phylodynamics of H5N1 avian influenza virus in Indonesia', *Molecular Ecology*, 21(12). doi: 10.1111/j.1365-294X.2012.05577.x.

Lam, T. T.-Y. *et al.* (2013) 'The genesis and source of the H7N9 influenza viruses causing human infections in China.', *Nature*. Nature Publishing Group, 502(7470), pp. 241–4. doi: 10.1038/nature12515.

Lam, T. T. Y. *et al.* (2013) 'Systematic phylogenetic analysis of influenza A virus reveals many novel mosaic genome segments', *Infection, Genetics and Evolution*. Elsevier B.V., 18, pp. 367–378. doi: 10.1016/j.meegid.2013.03.015.

Lamb, R., Krug, R. and Knipe, D. (2001) 'Orthomyxoviridae: The Viruses and Their Replication', in *Fields Virology*, pp. 1487–1531. doi: 10.1586/14787210.4.6.953.

Langat, P. *et al.* (2017) 'Genome-wide evolutionary dynamics of influenza B viruses on a global scale', *PLoS pathogens*, 13(12), p. e1006749. doi: 10.1371/journal.ppat.1006749.

Larison, B. *et al.* (2014) 'Spillover of pH1N1 to swine in Cameroon: an investigation of risk factors.', *BMC veterinary research*. BMC Veterinary Research, 10(1), p. 55. doi: 10.1186/1746-6148-10-55.

Leahy, M B *et al.* (1997) 'The fourth genus in the Orthomyxoviridae: sequence analyses of two Thogoto virus polymerase proteins and comparison with influenza viruses', *Virus research*, 50(2), pp. 215–224. doi: S0168-1702(97)00072-5 [pii].

Leahy, Michael B. *et al.* (1997) 'The fourth genus in the Orthomyxoviridae: Sequence analyses of two Thogoto virus polymerase proteins and comparison with influenza viruses', *Virus Research*, 50(2), pp. 215–224. doi: 10.1016/S0168-1702(97)00072-5.

Leamy, K. A. *et al.* (2016) 'Bridging the gap between in vitro and in vivo RNA folding', *Quarterly Reviews of Biophysics*, 49, p. e10. doi: 10.1017/S003358351600007X.

Lee, A. J. *et al.* (2015) 'Diversifying Selection Analysis Predicts Antigenic Evolution of 2009 Pandemic H1N1 Influenza A Virus in Humans', *Journal of Virology*, 89(10), pp. 5427–5440. doi: 10.1128/JVI.03636-14.

Lefkowitz, E. J. *et al.* (2017) 'Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)', *Nucleic Acids Research*. Oxford University Press, 1977(December), pp. 1–10. doi: 10.1093/nar/gkx932.

Lemey, P *et al.* (2009) 'Bayesian phylogeography finds its roots', *PLoS Comp. Biol.* , 5, p. e1000520. Available at: http://dx.doi.org/10.1371/journal.pcbi.1000520.

Lemey, Philippe *et al.* (2009) 'Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning.', *BMC bioinformatics*, 10, p. 126. doi: 10.1186/1471-2105-10-126.

Lemey, P. *et al.* (2010) 'Phylogeography takes a relaxed random walk in continuous space and time', *Mol. Biol. Evol.* , 27(8), pp. 1877–1885. doi: 10.1093/molbev/msq067.

Lemey, P. (2012) 'Phylogeographic inference in continuous space', (September), pp. 1–18.

Lemey, P. *et al.* (2012) 'The seasonal flight of influenza: a unified framework for spatiotemporal hypothesis testing', pp. 1–16. Available at: http://arxiv.org/abs/1210.5877.

Lemey, P. *et al.* (2014) 'Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2.', *PLoS pathogens*. Public Library of Science, 10(2), pp. 1–17. doi: 10.1371/journal.ppat.1003932.

Lemey, P., Rambaut, A. and Pybus, O. G. (2006) 'HIV evolutionary dynamics within and among hosts', *AIDS Reviews*, 8(3), pp. 125–140.

Lepage, T. *et al.* (2007) 'A general comparison of relaxed molecular clock models', *Molecular Biology and Evolution*, 24(Jeffreys 1935), pp. 2669–2680. doi: 10.1093/molbev/msm193.

Letunic, I. and Bork, P. (2019) 'Interactive Tree Of Life (iTOL) v4: recent updates and new developments', *Nucleic acids research*. Oxford University Press, 47(W1), pp. W256–W259. doi: 10.1093/nar/gkz239.

Leung, T. F. *et al.* (2003) 'Severe acute respiratory syndrome (SARS) in children: epidemiology, presentation and management.', *Paediatric respiratory reviews*, 4(4), pp. 334–339. doi: 10.1016/S1526.

Leventhal, G. *et al.* (2016) 'Epidemiology meets phylogenetics Inferring epidemiological dynamics based on genetic sequence data Epidemiology meets phylogenetics Inferring epidemiological dynamics based on genetic sequence data'.

Levy, B. and Odoi, A. (2018) 'Exploratory investigation of region level risk factors of ebola virus disease in West Africa', *PeerJ*, 2018(11), pp. 1–20. doi: 10.7717/peerj.5888.

Lewis, N. S. *et al.* (2015) 'Influenza A virus evolution and spatio-temporal dynamics in Eurasian Wild Birds: A phylogenetic and phylogeographic study of whole-genome sequence data.', *The Journal of general virology*, p. vir.0.000155-. doi: 10.1099/vir.0.000155.

Li, L. M., Grassly, N. C. and Fraser, C. (2014) 'Genomic analysis of emerging pathogens: methods, application and future trends', *Genome Biology*, 15(11), p. 541. doi: 10.1186/s13059-014-0541-9.

Li, W. (1993) 'Statistical Tests of Neutrality of Mutations'.

Li, W. *et al.* (2010) 'Genomic analysis of codon, sequence and structural conservation with selective biochemical-structure mapping reveals highly conserved and dynamic structures in rotavirus RNAs with potential cis-acting functions', *Nucleic Acids Research*, 38(21), pp. 7718–7735. doi: 10.1093/nar/gkq663.

Liang, L. *et al.* (2010) 'Combining spatial-temporal and phylogenetic analysis approaches for improved understanding on global H5N1 transmission.', *PloS one*, 5(10), p. e13575. doi: 10.1371/journal.pone.0013575.

Liang, L. J. *et al.* (2013) 'Evolution and mutation of H1N1pdm virus hemagglutinin genes in Guangdong', *Chinese Journal of Microbiology and Immunology (China)*, 33(3), pp. 173–177. doi: 10.3760/cma.j.issn.0254-5101.2013.03.003.

Libin, P. *et al.* (2017) 'PhyloGeoTool: Interactively exploring large phylogenies in an epidemiological context', *Bioinformatics*, 33(24), pp. 3993–3995. doi: 10.1093/bioinformatics/btx535.

Lin, J.-H. *et al.* (2011) 'Phylodynamics and molecular evolution of influenza A virus nucleoprotein genes in Taiwan between 1979 and 2009', *PLoS ONE*, 6(8). doi: 10.1371/journal.pone.0023454.

Lin, K. and Gallay, P. (2013) 'Curing a viral infection by targeting the host: the example of cyclophilin inhibitors.', *Antiviral research*. Elsevier B.V., 99(1), pp. 68–77. doi:

10.1016/j.antiviral.2013.03.020.

Lindstrom, S. E., Cox, N. J. and Klimov, A. (2004) 'Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events', *Virology*, 328(1), pp. 101–119. doi: 10.1016/j.virol.2004.06.009.

Liu, S.-Q. *et al.* (2015) 'Identifying the pattern of molecular evolution for Zaire ebolavirus in the 2014 outbreak in West Africa', *Infection, Genetics and Evolution*. Elsevier B.V., 32, pp. 51–59. doi: 10.1016/j.meegid.2015.02.024.

Liu, Z. Y. *et al.* (2016) 'Viral RNA switch mediates the dynamic control of flavivirus replicase recruitment by genome cyclization', *eLife*, 5(OCTOBER2016), pp. 1–27. doi: 10.7554/eLife.17636.

Lokody, I. (2014) 'RiboSNitches reveal heredity in RNA secondary structure', *Nature Reviews Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 15, p. 219. Available at: http://dx.doi.org/10.1038/nrg3700.

Lorenz, R. *et al.* (2016) 'Predicting RNA secondary structures from sequence and probing data', *Methods*. The Authors, 103, pp. 86–98. doi: 10.1016/j.ymeth.2016.04.004.

Lu, L., Lycett, S. J. and Brown, A. J. L. (2014) 'Reassortment patterns of avian influenza virus internal segments among different subtypes', *BMC evolutionary biology*, 14, p. 16. doi: 10.1186/1471-2148-14-16.

Ludwig, S. (2014) 'Will omics help to cure the flu?', *Trends in Microbiology*. Elsevier Ltd, 22(5), pp. 232–233. doi: 10.1016/j.tim.2014.03.003.

Lycett, S. J. *et al.* (2012) 'Estimating reassortment rates in co-circulating Eurasian swine influenza viruses', *Journal of General Virology*, 93(11), pp. 2326–2336. doi: 10.1099/vir.0.044503-0.

Magee, D. *et al.* (2014) 'Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion', *Archives of Virology*, 160(1), pp. 215–224. doi: 10.1007/s00705-014-2262-5.

Magee, D., Suchard, M. A. and Scotch, M. (2017) 'Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction', *PLoS Computational Biology*, 13(2), pp. 1–19. doi: 10.1371/journal.pcbi.1005389.

Maljkovic Berry, I. *et al.* (2016) 'Frequency of influenza H3N2 intra-subtype reassortment: attributes and implications of reassortant spread', *BMC Biology*. BMC Biology, 14(1), p. 117. doi: 10.1186/s12915-016-0337-3.

Mardones, F. O. *et al.* (2014) 'The role of fish movements and the spread of infectious salmon anemia virus (ISAV) in Chile, 2007-2009', *Preventive Veterinary Medicine*. Elsevier B.V., 114(1), pp. 37–46. doi: 10.1016/j.prevetmed.2014.01.012.

Markham, N. R. and Zuker, M. (2008) 'UNAFold: Software for nucleic acid folding and hybridization', *Methods in Molecular Biology*, 453, pp. 3–31. doi: 10.1007/978-1-60327-429-6-1.

Marra, M. A. (2003) 'The genome sequence of the SARS-associated coronavirus', *Science*. , 300, pp. 1399–1404. Available at: http://dx.doi.org/10.1126/science.1085953.

Marshall, N. *et al.* (2013) 'Influenza Virus Reassortment Occurs with High Frequency in the Absence of Segment Mismatch', 9(6), pp. 1–11. doi: 10.1371/journal.ppat.1003421.

Marston, H. D. *et al.* (2014) 'Emerging Viral Diseases: Confronting Threats with New Technologies.', *Science translational medicine*, 6(253), p. 253ps10. doi: 10.1126/scitranslmed.3009872.

Martín-Benito, J. and Ortín, J. (2013) 'Influenza Virus Transcription and Replication', *Advances in Virus Research*, 87, pp. 113–137. doi: 10.1016/B978-0-12-407698-3.00004-1.

Martin, D. P. *et al.* (2015) 'RDP4: Detection and analysis of recombination patterns in virus genomes', *Virus Evolution*, 1(1), pp. vev003–vev003. doi: 10.1093/ve/vev003.

Mason, J. (2016) 'Point-of-Care Testing for Influenza', *CADTH Issues in Emerging Health Technologies*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27977096.

Mathews, D. H. (2005) 'Predicting a set of minimal free energy RNA secondary structures common to two sequences', *Bioinformatics*, 21(10), pp. 2246–2253. doi: 10.1093/bioinformatics/bti349.

Matsuzaki, Y. *et al.* (2003) 'Frequent Reassortment among Influenza C Viruses', *Journal of Virology*, 77(2), pp. 871–881. doi: 10.1128/JVI.77.2.871-881.2003.

Matsuzaki, Y. *et al.* (2004) 'Genetic diversity of influenza B virus: The frequent reassortment and cocirculation of the genetically distinct reassortant viruses in a community', *Journal of Medical Virology*, 74(1), pp. 132–140. doi: 10.1002/jmv.20156.

Matsuzaki, Y. *et al.* (2016) 'Genetic lineage and reassortment of influenza C viruses circulating between 1947 and 2014.', *Journal of virology*, (July). doi: 10.1128/JVI.00969-16.

Mayer, S. V., Tesh, R. B. and Vasilakis, N. (2017) 'The emergence of arthropod-borne viral diseases: A global prospective on dengue, chikungunya and zika fevers', *Acta Tropica*. Elsevier B.V., 166, pp. 155–163. doi: 10.1016/j.actatropica.2016.11.020.

McDonald, S. M. *et al.* (2016) 'Reassortment in segmented RNA viruses: mechanisms and outcomes', *Nature Reviews Microbiology*. Nature Publishing Group, (May). doi: 10.1038/nrmicro.2016.46.

McGill, J. R., Walkup, E. A. and Kuhner, M. K. (2013) 'Graphml specializations to codify ancestral recombinant graphs', *Frontiers in Genetics*, 4(AUG), pp. 1–5. doi: 10.3389/fgene.2013.00146.

Mehle, A. *et al.* (2012) 'Reassortment and Mutation of the Avian Influenza Virus Polymerase PA Subunit Overcome Species Barriers', *Journal of Virology*, 86(3), pp. 1750–1757. doi: 10.1128/jvi.06203-11.

Merler, S. *et al.* (2011) 'Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: Implications for real-time modelling', *PLoS Computational Biology*, 7(9). doi: 10.1371/journal.pcbi.1002205.

Mérour, E. *et al.* (2011) 'Completion of the full-length genome sequence of the infectious salmon anemia virus, an aquatic orthomyxovirus-like, and characterization of mAbs', *Journal of General Virology*, 92(3), pp. 528–533. doi: 10.1099/vir.0.027417-0.

Metcalf, C. J. E. *et al.* (2015) 'Five challenges in evolution and infectious diseases', *Epidemics*. Elsevier B.V., 10, pp. 40–44. doi: 10.1016/j.epidem.2014.12.003.

Metzker, M. L. (2010) ‘Sequencing technologies - the next generation.’, *Nature reviews. Genetics*, 11(1), pp. 31–46.

Meyer, A. G. *et al.* (2015) ‘Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak’, *Virus Evolution*, 1(1), p. vev006. doi: 10.1093/ve/vev006.

Miller, M. A. *et al.* (2009) ‘The Signature Features of Influenza Pandemics — Implications for Policy’, *New England Journal of Medicine*, 360(25), pp. 2595–2598. doi: 10.1056/NEJMp0903906.

Mjaaland, S. *et al.* (1997) ‘Genomic characterization of the virus causing infectious salmon anemia in Atlantic salmon (Salmo salar L.): an orthomyxo-like virus in a teleost.’, *Journal of virology*, 71(10), pp. 7681–6. Available at: http://jvi.asm.org/cgi/content/abstract/71/10/7681%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=192118&tool=pmcentrez&rendertype=abstract.

Monamele, G. C. *et al.* (2017) ‘Associations between meteorological parameters and influenza activity in a subtropical country: Case of five sentinel sites in Yaoundé-Cameroon’, *PLoS ONE*, 12(10), pp. 1–12. doi: 10.1371/journal.pone.0186914.

Mondini, A. *et al.* (2009) ‘Spatio-temporal tracking and phylodynamics of an urban dengue 3 outbreak in São Paulo, Brazil.’, *PLoS neglected tropical diseases*, 3(5).

Morens, D. M., Taubenberger, J. K. and Fauci, a. S. (2010) ‘The 2009 H1N1 Pandemic Influenza Virus: What Next?’, *mBio*, 1(4), p. e00211. doi: 10.1128/mBio.00211-10.Updated.

Morens, D. M., Taubenberger, J. K. and Fauci, A. S. (2009) ‘The persistent legacy of the 1918 influenza virus.’, *The New England journal of medicine*, 361(3), pp. 225–229. doi: 10.1056/NEJMp0904819.

Morse, S. S. *et al.* (2012) ‘Prediction and prevention of the next pandemic zoonosis’, *The Lancet*. Elsevier Ltd, 380(9857), pp. 1956–1965. doi: 10.1016/S0140-6736(12)61684-5.

Moss, W. N. *et al.* (2012) ‘The influenza A segment 7 mRNA 3’ splice site pseudoknot/hairpin family’, *RNA Biol*, 9(11), pp. 1305–1310. doi: 10.4161/rna.22343.

Moss, W. N., Priore, S. F. and Turner, D. H. (2011) ‘Identification of potential conserved RNA secondary structure throughout influenza A coding regions.’, *RNA (New York, N.Y.)*, 17(6), pp. 991–1011. doi: 10.1261/rna.2619511.

El Moussi, A. *et al.* (2013) ‘Virological Surveillance of Influenza Viruses during the 2008-09, 2009-10 and 2010-11 Seasons in Tunisia.’, *PloS one*, 8(9), p. e74064. doi: 10.1371/journal.pone.0074064.

Muhire, B. M. *et al.* (2014) ‘Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses.’, *Journal of virology*, 88(4), pp. 1972–89. doi: 10.1128/JVI.03031-13.

Munshili Njifon, H. L. *et al.* (2018) ‘Influence of meteorological parameters in the seasonality of influenza viruses circulating in Northern Cameroon’, *Influenza and other Respiratory Viruses*, (March 2018), pp. 158–165. doi: 10.1111/irv.12612.

Murakami, S. *et al.* (2016) ‘Influenza d virus infection in herd of cattle, Japan’, *Emerging Infectious Diseases*, 22(8), pp. 1517–1519. doi: 10.3201/eid2208.160362.

Murrell, B. *et al.* (2013) 'FUBAR : A Fast , Unconstrained Bayesian AppRoximation for Inferring Selection', 30(5), pp. 1196–1205. doi: 10.1093/molbev/mst030.

Nagarajan, N. and Kingsford, C. (2011) 'GiRaF: robust, computational identification of influenza reassortments via graph mining.', *Nucleic acids research*, 39(6), p. e34. doi: 10.1093/nar/gkq1232.

Nascimento, F. F., Reis, M. Dos and Yang, Z. (2017) 'A biologist's guide to Bayesian phylogenetic analysis', *Nature Ecology and Evolution*. Springer US, 1(10), pp. 1446–1454. doi: 10.1038/s41559-017-0280-x.

Neatherlin, J. *et al.* (2013) 'Influenza A(H1N1)pdm09 during air travel.', *Travel medicine and infectious disease*, 11(2), pp. 110–8. doi: 10.1016/j.tmaid.2013.02.004.

Neher, R. a. *et al.* (2015) 'Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses', *arXiv*, pp. 1–13. doi: 10.1073/pnas.1525578113.

Neher, R. A., Russell, C. A. and Shraiman, B. I. (2014) 'Predicting evolution from the shape of genealogical trees.', *eLife*, 3, p. e03568. doi: 10.7554/eLife.03568.

Nelson, M. I. *et al.* (2008) 'Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918', *PLoS Pathogens*, 4(2). doi: 10.1371/journal.ppat.1000012.

Nelson, M. I. *et al.* (2012) 'Global transmission of influenza viruses from humans to swine', *Journal of General Virology*, 93(PART 10), pp. 2195–2203. doi: 10.1099/vir.0.044974-0.

Nelson, M. I. *et al.* (2014) 'Multiyear persistence of 2 pandemic A/H1N1 influenza virus lineages in West Africa', *Journal of Infectious Diseases*, 210(1), pp. 121–125. doi: 10.1093/infdis/jiu047.

Nelson, M. I. *et al.* (2015) 'Continual Reintroduction of Human Pandemic H1N1 Influenza A Viruses into Swine in the United States, 2009 to 2014', *Journal of Virology*, 89(12), pp. 6218–6226. doi: 10.1128/jvi.00459-15.

Nelson, M. I. and Holmes, E. C. (2007) 'The evolution of epidemic influenza', *Nature Reviews Genetics*, 8(3), pp. 196–205. doi: 10.1038/nrg2053.

Neumann, G. and Kawaoka, Y. (2015) 'Transmission of in fl uenza A viruses', *Virology*, 480, pp. 234–246. doi: 10.1016/j.virol.2015.03.009.

Neumann, G., Noda, T. and Kawaoka, Y. (2009) 'Emergence and pandemic potential of swine-origin H1N1 influenza virus.', *Nature*, 459(7249), pp. 931–9. doi: 10.1038/nature08157.

Neverov, A. D. *et al.* (2014) 'Intrasubtype Reassortments Cause Adaptive Amino Acid Replacements in H3N2 Influenza Genes', *PLoS Genetics*, 10(1), pp. 12–14. doi: 10.1371/journal.pgen.1004037.

Nicolaides, C. *et al.* (2012) 'A metric of influential spreading during contagion dynamics through the air transportation network.', *PloS one*, 7(7), p. e40961. doi: 10.1371/journal.pone.0040961.

Noble, W. S. (2009) 'How does multiple testing correction work?', *Nature Biotechnology*. Nature Publishing Group, 27(12), pp. 1135–1137. doi: 10.1038/nbt1209-1135.

Norström, M. M., Karlsson, A. C. and Salemi, M. (2012) 'Towards a new paradigm linking

virus molecular evolution and pathogenesis: Experimental design and phylodynamic inference', *New Microbiologica*, 35(2).

Nunes, M. R. T. *et al.* (2012) 'Phylogeography of dengue virus serotype 4, Brazil, 2010-2011', *Emerging Infectious Diseases*, 18(11), pp. 1858–1864. doi: 10.3201/eid1811.120217.

Nunes, M. R. T. *et al.* (2014) 'Air travel is associated with intracontinental spread of dengue virus serotypes 1-3 in Brazil.', *PLoS neglected tropical diseases*. Public Library of Science, 8(4), p. e2769. doi: 10.1371/journal.pntd.0002769.

Nzussouo, N. T. *et al.* (2012) 'Delayed 2009 pandemic influenza A virus subtype H1N1 circulation in West Africa, May 2009-April 2010.', *The Journal of infectious diseases*, 206 Suppl(November 2009), pp. S101-7. doi: 10.1093/infdis/jis572.

Odagiri, T. *et al.* (2015) 'Isolation and characterization of influenza C viruses in the Philippines and Japan', *Journal of Clinical Microbiology*, 53(3), pp. 847–858. doi: 10.1128/JCM.02628-14.

Ojosnegros, S. *et al.* (2011) 'Viral genome segmentation can result from a trade-off between genetic content and particle stability', *PLoS Genetics*, 7(3), pp. 1–11. doi: 10.1371/journal.pgen.1001344.

Olmstead, A. D. *et al.* (2015) 'A molecular phylogenetics-based approach for identifying recent hepatitis C virus transmission events', *Infection, Genetics and Evolution*. Elsevier B.V., 33, pp. 101–109. doi: 10.1016/j.meegid.2015.04.017.

Oong, X. Y. *et al.* (2015) 'Epidemiological and evolutionary dynamics of influenza B viruses in Malaysia, 2012-2014', *PLoS ONE*, 10(8), pp. 2012–2014. doi: 10.1371/journal.pone.0136254.

Ortiz, J. R. *et al.* (2012) 'Pandemic influenza in Africa, lessons learned from 1968: A systematic review of the literature', *Influenza and other Respiratory Viruses*, 6(1), pp. 11–24. doi: 10.1111/j.1750-2659.2011.00257.x.

Owoade, A. A. *et al.* (2008) 'Replacement of sublineages of avian influenza (H5N1) by reassortments, sub-Saharan Africa', *Emerging Infectious Diseases*, 14(11), pp. 1731–1735. doi: 10.3201/eid1411.080555.

Pachler, K. and Vlasak, R. (no date) 'Whole Genome Sequencing of Influenza C Virus'.

Pagel, M., Meade, A. and Barker, D. (2004) 'Bayesian Estimation of Ancestral Character States on Phylogenies', *Systematic Biology*, 53(5), pp. 673–684. doi: 10.1080/10635150490522232.

Paradis, E. and Schliep, K. (2019) 'Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics*, 35(3), pp. 526–528. doi: 10.1093/bioinformatics/bty633.

Paraskevis, D. *et al.* (2015) 'Enhanced HIV-1 surveillance using molecular epidemiology to study and monitor HIV-1 outbreaks among intravenous drug users (IDUs) in Athens and Bucharest', *Infection, Genetics and Evolution*, 35. doi: 10.1016/j.meegid.2015.08.004.

Pedersen, J. S. *et al.* (2004) 'A comparative method for finding and folding RNA secondary structures within protein-coding regions', *Nucleic Acids Research*, 32(16), pp. 4925–4936. doi: 10.1093/nar/gkh839.

Peng, G. *et al.* (1994) 'Genetic reassortment of influenza C viruses in man', *Journal of General Virology*, 75(12), pp. 3619–3622. doi: 10.1099/0022-1317-75-12-3619.

Peng, R. *et al.* (2017) 'Structures of human-infecting *Thogotovirus* fusogens support a common ancestor with insect baculovirus', *Proceedings of the National Academy of Sciences*, 114(42), pp. E8905–E8912. doi: 10.1073/pnas.1706125114.

Plarre, H. *et al.* (2012) 'Evolution of infectious salmon anaemia virus (ISA virus)', *Archives of Virology*, 157(12), pp. 2309–2326. doi: 10.1007/s00705-012-1438-0.

du Plessis, L. and Stadler, T. (2015) 'Getting to the root of epidemic spread with phylodynamic analysis of genomic data', *Trends in Microbiology*. Elsevier Ltd, 23(7), pp. 383–386. doi: 10.1016/j.tim.2015.04.007.

Pompei, S., Loreto, V. and Tria, F. (2012) 'Phylogenetic Properties of RNA Viruses', *PLoS ONE*, 7(9). doi: 10.1371/journal.pone.0044849.

Pond, S. L. K., Poon, A. and Frost, S. D. W. (2009) 'Estimating selection pressures on alignments of coding sequences', *—Lemey P, Salemi M, Vandamme A*, pp. 1–81. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Estimating+selection+pressures+on+alignments+of+coding+sequences#4.

Poon, A. F. Y., Frost, S. D. W. and Pond, S. L. K. (2009) 'Detecting signatures of selection from DNA sequences using datamonkey', *Methods in Molecular Biology*, 537, pp. 163–183. doi: 10.1007/978-1-59745-251-9_8.

Poon, L. L. M. *et al.* (2016) 'Quantifying influenza virus diversity and transmission in humans', *Nature Genetics*. Nature Publishing Group, 48(2), pp. 195–200. doi: 10.1038/ng.3479.

Posada, D. (2008) 'jModelTest: Phylogenetic Model Averaging', *Molecular Biology and Evolution*, 25(7), pp. 1253–1256. doi: 10.1093/molbev/msn083.

Presti, R. M. *et al.* (2009) 'Quaranfil, Johnston Atoll, and Lake Chad Viruses Are Novel Members of the Family Orthomyxoviridae', *Journal of Virology*, 83(22), pp. 11599–11606. doi: 10.1128/JVI.00677-09.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010) 'FastTree 2 - Approximately maximum-likelihood trees for large alignments', *PLoS ONE*, 5(3). doi: 10.1371/journal.pone.0009490.

Priore, S. F. *et al.* (2013) 'Secondary structure of a conserved domain in the intron of influenza A NS1 mRNA.', *PloS one*, 8(9). Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84892832702&partnerID=tZOtx3y1.

Priore, S. F. *et al.* (2015) 'The Influenza A PB1-F2 & N40 Start Codons are Contained Within an RNA Pseudoknot', *Biochemistry*. American Chemical Society, 54(22), p. 150521120657006. doi: 10.1021/bi501564d.

Priore, S. F., Moss, W. N. and Turner, D. H. (2012) 'Influenza A Virus Coding Regions Exhibit Host-Specific Global Ordered RNA Structure', *PLoS ONE*, 7(4), p. e35989. doi: 10.1371/journal.pone.0035989.

Priore, S. F., Moss, W. N. and Turner, D. H. (2013) 'Influenza B virus has global ordered RNA structure in (+) and (-) strands but relatively less stable predicted RNA folding free

energy than allowed by the encoded protein sequence.', *BMC research notes*, 6(1), p. 330. doi: 10.1186/1756-0500-6-330.

Prosperi, M. *et al.* (2013) 'Molecular Epidemiology of Community- Staphylococcus aureus in the genomic era : a Cross-Sectional Study', pp. 1–8. doi: 10.1038/srep01902.

Puzelli, S. *et al.* (2004) 'Changes in the hemagglutinins and neuraminidases of human influenza B viruses isolated in Italy during the 2001-02, 2002-03, and 2003-04 seasons', *Journal of Medical Virology*, 74(4), pp. 629–640. doi: 10.1002/jmv.20225.

Pybus, O. G. and Rambaut, A. (2009) 'Evolutionary analysis of the dynamics of viral infectious disease', *Nature Reviews Genetics*, 10(8), pp. 540–550. doi: 10.1038/nrg2583.

Radin, J. M., Katz, M. A., Tempia, S., Nzussouo, N. T., *et al.* (2012) 'In fl uenza Surveillance in 15 Countries in Africa , 2006 – 2010', 206(Suppl 1), pp. 2006–2010. doi: 10.1093/infdis/jis606.

Radin, J. M., Katz, M. A., Tempia, S., Talla Nzussouo, N., *et al.* (2012) 'Influenza surveillance in 15 countries in Africa, 2006-2010.', *The Journal of infectious diseases*, 206 Suppl(suppl_1), pp. S14-21. doi: 10.1093/infdis/jis606.

Rahnama, L. and Aris-Brosou, S. (2013) 'Phylodynamics of the emergence of influenza viruses after cross-species transmission.', *PloS one*, 8(12), p. e82486. doi: 10.1371/journal.pone.0082486.

Rajao, D. S. *et al.* (2014) *Evolution and Ecology of Influenza A Viruses, Current Topics in Microbiology and immunology Influenza Pathogenesis and Control-Volume I*. doi: 10.1007/978-3-319-1115-1.

Rambaut, A. *et al.* (2016) 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2(1), p. vew007. doi: 10.1093/ve/vew007.

Rambaut, A., Drummond, A. J. and Suchard, M. (2003) *Tracer v1.6: MCMC Trace Analysis Package, [computer program]*. Available at: http://tree.bio.ed.ac.uk/software/tracer/.

Ramey, A. M. *et al.* (2010) 'Intercontinental reassortment and genomic variation of low pathogenic avian influenza viruses isolated from northern pintails (Anas acuta) in Alaska: Examining the evidence through space and time', *Virology*. Elsevier B.V., 401(2), pp. 179–189. doi: 10.1016/j.virol.2010.02.006.

Rao, D. M. (2009) 'Enhancing Temporo-Geospatial Epidemiological Analysis of H5N1 Influenza using Phylogeography', pp. 1–12.

Rasmussen, D. A., Ratmann, O. and Koelle, K. (2011) 'Inference for nonlinear epidemiological models using genealogies and time series', *PLoS Computational Biology*, 7(8). doi: 10.1371/journal.pcbi.1002136.

Ratmann, O. *et al.* (2017) 'Phylogenetic tools for generalized HIV-1 epidemics: Findings from the PANGEA-HIV methods comparison', *Molecular Biology and Evolution*, 34(1), pp. 185–203. doi: 10.1093/molbev/msw217.

Re, S. (2005) 'Influenza Pandemics', *Vaccine*, p. S15. doi: 10.1016/S0264-410X(02)00122-6.

Redelings, B. D. and Suchard, M. A. (2005) 'Joint Bayesian estimation of alignment and phylogeny', *Syst. Biol.* , 54, pp. 401–418. Available at:

http://dx.doi.org/10.1080/10635150590947041.

Reid, A. H. *et al.* (2002) 'Characterization of the 1918 "Spanish" influenza virus matrix gene segment.', *Journal of virology*, 76(21), pp. 10717–10723. doi: 10.1128/JVI.76.21.10717-10723.2002.

Reperant, L. A., Kuiken, T. and Osterhaus, A. D. M. E. (2012) 'Adaptive pathways of zoonotic influenza viruses : From exposure to establishment in humans', *Vaccine*. Elsevier Ltd, 30(30), pp. 4419–4434. doi: 10.1016/j.vaccine.2012.04.049.

Resa-Infante, P. *et al.* (2011) 'The influenza virus RNA synthesis machine: advances in its structure and function', *RNA Biol*, 8(2), pp. 207–215. doi: 10.4161/hv.8.2.14513.

Rieppel, O. (2011) 'Ernst Haeckel (1834-1919) and the monophyly of life', *Journal of Zoological Systematics and Evolutionary Research*. doi: 10.1111/j.1439-0469.2010.00580.x.

Rife, B. D. *et al.* (2017) 'Phylodynamic applications in 21st century global infectious disease research', *Global Health Research and Policy*. Global Health Research and Policy, 2(1), p. 13. doi: 10.1186/s41256-017-0034-y.

Rivas, a L. *et al.* (2010) 'Lessons from Nigeria: the role of roads in the geo-temporal progression of avian influenza (H5N1) virus.', *Epidemiology and infection*, 138(2), pp. 192–8. doi: 10.1017/S0950268809990495.

Rota, P. A. *et al.* (1990) 'Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983', *Virology*, 175(1), pp. 59–68. doi: 10.1016/0042-6822(90)90186-U.

Saéz, A. M. *et al.* (2014) 'Investigating the zoonotic origin of the West African Ebola epidemic', *EMBO Molecular Medicine*, 7(1), pp. 17–23. doi: 10.15252/emmm.201404792.

Sander, B. *et al.* (2010) 'Is a mass immunization program for pandemic (H1N1) 2009 good value for money? Evidence from the Canadian Experience', *Vaccine*, 28(38), pp. 6210–6220. doi: 10.1016/j.vaccine.2010.07.010.

Santiago, F. V. and Rivera-Amill, V. (2015) 'Envelope Gene Evolution and HIV-1 Neuropathogenesis', *Journal of Neuroinfectious Diseases*, s2(Suppl 2), pp. 1–15. doi: 10.4172/2314-7326.s2-003.

Schierup, M. H. and Hein, J. (2000) 'Consequences of recombination on traditional phylogenetic analysis', *Genetics*, 156(2), pp. 879–891. doi: 10.1668/0003-1569(2001)041[0134:BR]2.0.CO;2.

Scholtissek, C. *et al.* (1978a) 'On the origin of the human influenza virus subtypes H2N2 and H3N2', *Virology*. Academic Press, 87(1), pp. 13–20. doi: 10.1016/0042-6822(78)90153-8.

Scholtissek, C. *et al.* (1978b) 'On the origin of the human influenza virus subtypes H2N2 and H3N2', *Virology*, 87(1), pp. 13–20. doi: 10.1016/0042-6822(78)90153-8.

Scholtissek, C. (1994) 'Source for influenza pandemics', *European Journal of Epidemiology*, 10(4), pp. 455–458. doi: 10.1007/BF01719674.

Schoub, Barry D *et al.* (2013) 'Afriflu2--second international workshop on influenza vaccination in the African continent--8 November 2012, Cape Town (South Africa).', *Vaccine*, 31(35), pp. 3461–6. doi: 10.1016/j.vaccine.2013.04.021.

Schoub, Barry D. *et al.* (2013) 'Afriflu2-Second international workshop on influenza

vaccination in the African continent-8 November 2012, Cape Town (South Africa)', *Vaccine*, 31(35), pp. 3461–3466. doi: 10.1016/j.vaccine.2013.04.021.

Schrauwen, E. J. and Fouchier, R. A. (2014) 'Host adaptation and transmission of influenza A viruses in mammals', *Emerging Microbes & Infections*, 3(2), p. e9. doi: 10.1038/emi.2014.9.

Seetin, M. G. and Mathews, D. H. (2012) 'RNA Structure Prediction: An Overview of Methods', in Keiler, K. C. (ed.) *Bacterial Regulatory RNA: Methods and Protocols*. Totowa, NJ: Humana Press, pp. 99–122. doi: 10.1007/978-1-61779-949-5_8.

Semegni, J. Y. *et al.* (2011) 'NASP: A parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments', *Bioinformatics*, 27(17), pp. 2443–2445. doi: 10.1093/bioinformatics/btr417.

Sessions, O. M. *et al.* (2013) 'Exploring the origin and potential for spread of the 2013 dengue outbreak in Luanda, Angola.', *Global health action*, 6(1), p. 21822. doi: 10.3402/gha.v6i0.21822.

Shapiro, B., Rambaut, A. and Drummond, A. J. (2006) 'Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences', *Mol. Biol. Evol.* , 23, pp. 7–9. Available at: http://dx.doi.org/10.1093/molbev/msj021.

Shi, W. *et al.* (2010) 'A complete analysis of HA and NA genes of influenza A viruses.', *PloS one*, 5(12), p. e14454. doi: 10.1371/journal.pone.0014454.

de Silva, E., Ferguson, N. M. and Fraser, C. (2012) 'Inferring pandemic growth rates from sequence data.', *Journal of the Royal Society, Interface / the Royal Society*, 9(73), pp. 1797–808. doi: 10.1098/rsif.2011.0850.

Sim, S. and Hibberd, M. L. (2016) 'Genomic approaches for understanding dengue: insights from the virus, vector, and host', *Genome Biology*. Genome Biology, 17(1), p. 38. doi: 10.1186/s13059-016-0907-2.

Simon-Loriere, E. and Holmes, E. C. (2011) 'Why do RNA viruses recombine?', *Nature Reviews Microbiology*. Nature Publishing Group, 9(8), pp. 617–626. doi: 10.1038/nrmicro2614.

Simon, A. E. and Gehrke, L. (2009) 'RNA conformational changes in the life cycles of RNA viruses, viroids, and virus-associated RNAs', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Elsevier B.V., 1789(9–10), pp. 571–583. doi: 10.1016/j.bbagrm.2009.05.005.

Sintchenko, V. and Holmes, E. C. (2015) 'The role of pathogen genomics in assessing disease transmission', *Bmj*, 350(may11 1), pp. h1314–h1314. doi: 10.1136/bmj.h1314.

Smith, G. J. D., Bahl, J., *et al.* (2009) 'Dating the emergence of pandemic influenza viruses.', *Proceedings of the National Academy of Sciences of the United States of America*, 106(28), pp. 11709–12. doi: 10.1073/pnas.0904991106.

Smith, G. J. D., Vijaykrishna, D., *et al.* (2009) 'Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.', *Nature*, 459(7250), pp. 1122–5. doi: 10.1038/nature08182.

Snoeck O. J.; Sausy, A.; Okwen, M. P.; Olubayo, A. G.; Owoade, A. A.; Muller, C. P., C. J. .

A. (2015) ‘Serological evidence of pandemic (H1N1) 2009 virus in pigs, West and Central Africa’, *Veterinary Microbiology*, 176(1–2), pp. 165–171. doi: 10.1016/j.vetmic.2014.12.022.

Soszynska-Jozwiak, M. *et al.* (2017) ‘Influenza virus segment 5 (+)RNA - Secondary structure and new targets for antiviral strategies’, *Scientific Reports*. Springer US, 7(1), pp. 1–15. doi: 10.1038/s41598-017-15317-5.

Speranskaya, A. S. *et al.* (2012) ‘Genetic diversity and evolution of the influenza C virus’, *Russian Journal of Genetics*, 48(7), pp. 671–678. doi: 10.1134/S1022795412070149.

Spielman, S. J. and Wilke, C. O. (2015) ‘The Relationship between dN/dS and Scaled Selection Coefficients.’, *Molecular biology and evolution*, 32(4), pp. 1097–1108. doi: 10.1093/molbev/msv003.

Spirollari, J. *et al.* (2009) ‘Predicting consensus structures for RNA alignments via pseudo-energy minimization’, *Bioinformatics and Biology Insights*, 2009(3), pp. 51–69.

Stack, J. C. *et al.* (2010) ‘Protocols for sampling viral sequences to study epidemic dynamics.’, *Journal of the Royal Society, Interface / the Royal Society*, 7(48), pp. 1119–1127. doi: 10.1098/rsif.2009.0530.

Stauffer, R. C. (2004) ‘Haeckel, Darwin, and Ecology’, *The Quarterly Review of Biology*. doi: 10.1086/401754.

Steel, J. and Lowen, A. C. (2014) ‘Influenza A Virus Reassortment’, (July), pp. 377–401. doi: 10.1007/82.

Steffen, C. *et al.* (2011) ‘Afriflu--international conference on influenza disease burden in Africa, 1-2 June 2010, Marrakech, Morocco.’, *Vaccine*, 29(3), pp. 363–9. doi: 10.1016/j.vaccine.2010.11.029.

Studies, I. P. (2008) ‘Spatio-temporal patterns in the transmissibility of influenza’, *Lancet, The*.

Su, Y. C. F. *et al.* (2015a) ‘Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection.’, *Nature communications*. Nature Publishing Group, 6, p. 7952. doi: 10.1038/ncomms8952.

Su, Y. C. F. *et al.* (2015b) ‘Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection.’, *Nature communications*. Nature Publishing Group, 6, p. 7952. doi: 10.1038/ncomms8952.

Suchard, M. A. *et al.* (2018) ‘Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10’, *Virus Evolution*, 4(1), pp. 1–5. doi: 10.1093/ve/vey016.

Suchard, M. A. and Drummond, A. J. (2010) ‘Bayesian random local clocks, or one rate to rule them all’, *BMC Biology*, 8(1), p. 114. Available at: http://www.biomedcentral.com/1741-7007/8/114.

Sun, H. *et al.* (2013) ‘Using sequence data to infer the antigenicity of influenza virus’, *mBio*, 4(4), pp. 1–9. doi: 10.1128/mBio.00230-13.

Suptawiwat, O. *et al.* (2017) ‘Evolutionary dynamic of antigenic residues on influenza B hemagglutinin’, *Virology*. Elsevier, 502(December 2016), pp. 84–96. doi: 10.1016/j.virol.2016.12.015.

Szewczyk, B., Bienkowska-Szewczyk, K. and Król, E. (2014) 'Introduction to molecular biology of influenza A viruses', *Acta Biochimica Polonica*, 61(3), pp. 397–401. doi: Epub 2014 Sep 3.

Tajima, F. (1989) 'Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism', 595(3), pp. 585–595.

Takebe, Y. *et al.* (2010) 'Reconstructing the epidemic history of HIV-1 circulating recombinant forms CRF07_BC and CRF08_BC in East Asia: The relevance of genetic diversity and phylodynamics for vaccine strategies', *Vaccine*, 28(SUPPL. 2). doi: 10.1016/j.vaccine.2009.07.101.

Talbi, C. *et al.* (2010) 'Phylodynamics and Human-mediated dispersal of a zoonotic virus', *PLoS Pathogens*, 6(10). doi: 10.1371/journal.ppat.1001166.

Talla Nzussouo, N. *et al.* (2017) 'Epidemiology of influenza in West Africa after the 2009 influenza A(H1N1) pandemic, 2010-2012', *BMC Infectious Diseases*. BMC Infectious Diseases, 17(1), pp. 2010–2012. doi: 10.1186/s12879-017-2839-1.

Tao, H., Steel, J. and Lowen, A. C. (2014) 'Intrahost Dynamics of Influenza Virus Reassortment', *Journal of Virology*, 88(13), pp. 7485–7492. doi: 10.1128/JVI.00715-14.

Tatem, A. and Rogers, D. (2006) 'Global transport networks and infectious disease spread', *Advances in parasitology*, 62(05), pp. 293–343. doi: 10.1016/S0065-308X(05)62009-X.Global.

Taubenberger, J. *et al.* (2005) 'Characterization of the 1918 influenza virus polymerase genes.', *Nature*, 437(7060), pp. 889–93. doi: 10.1038/nature04230.

Taubenberger, J. K. *et al.* (1997) 'Initial genetic characterization of the 1918 "Spanish" influenza virus.', *Science (New York, N.Y.)*, 275(5307), pp. 1793–1796. doi: 10.1126/science.275.5307.1793.

Taubenberger, J. K. *et al.* (2001) 'Integrating historical, clinical and molecular genetic data in order to explain the origin and virulence of the 1918 Spanish influenza virus.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 356(1416), pp. 1829–1839. doi: 10.1098/rstb.2001.1020.

Taubenberger, J. K. and Morens, D. M. (2006) '1918 Influenza: The mother of all pandemics', *Emerging Infectious Diseases*, 12(1), pp. 15–22. doi: 10.3201/eid1201.050979.

Taubenberger, J. K. and Morens, D. M. (2013) 'Influenza viruses: Breaking all the rules', *mBio*, 4(4), pp. 1–6. doi: 10.1128/mBio.00365-13.

Tee, K. K. *et al.* (2009) 'Estimating the date of origin of an HIV-1 circulating recombinant form', *Virology*. Elsevier Inc., 387(1), pp. 229–234. doi: 10.1016/j.virol.2009.02.020.

Tempia, S. *et al.* (2015) 'Mortality Associated with Seasonal and Pandemic Influenza among Pregnant and Non-Pregnant Women of Childbearing Age in a High HIV Prevalence Setting - South Africa, 1999-2009.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 61, pp. 1063–1070. doi: 10.1093/cid/civ448.

Tewawong, N. *et al.* (2015) 'Assessing antigenic drift of seasonal influenza A(H3N2) and A(H1N1)pdm09 viruses', *PLoS ONE*, 10(10), pp. 1–15. doi: 10.1371/journal.pone.0139958.

Than, V. T., Baek, I. H. and Kim, W. (2013) 'Whole genomic analysis reveals the co-

evolutionary phylodynamics of Korean G9P[8] human rotavirus strains', *Archives of Virology*, 158(8). doi: 10.1007/s00705-013-1662-2.

Theo, A. *et al.* (2012) 'Influenza surveillance in Zambia, 2008-2009.', *The Journal of infectious diseases*, 206 Suppl(Suppl 1), pp. S173-7. doi: 10.1093/infdis/jis599.

Theys, K. *et al.* (2019) 'Advances in Visualization Tools for Phylogenomic and Phylodynamic Studies of Viral Diseases', *Frontiers in Public Health*, 7(August), pp. 1–18. doi: 10.3389/fpubh.2019.00208.

Tizzoni, M. *et al.* (2012) 'Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm.', *BMC medicine*. BioMed Central Ltd, 10(1), p. 165. doi: 10.1186/1741-7015-10-165.

Tongo, M. *et al.* (2018) 'Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages', *Virus Evolution*, 4(1), pp. 1–13. doi: 10.1093/ve/vey003.

Tramuto, F. *et al.* (2016) 'The molecular epidemiology and evolutionary dynamics of influenza B virus in two Italian regions during 2010-2015: The experience of Sicily and Liguria', *International Journal of Molecular Sciences*, 17(4), pp. 1–15. doi: 10.3390/ijms17040549.

Trifonov, E. N. (1997) 'Segmented Structure of Separate and Transposable DNA and RNA Elements as Suggested by Their Size Distributions', *Journal of Biomolecular Structure & Dynamics*, 14(4), pp. 449–457. doi: 10.1080/07391102.1997.10508144.

Trifonov, V. *et al.* (2009) 'The origin of the recent swine influenza A(H1N1) virus infecting humans.', *Euro surveillance : bulletin europ??en sur les maladies transmissibles = European communicable disease bulletin*, 14(17), p. 19193. doi: 19193 [pii].

Trifonov, V., Khiabanian, H. and Rabadan, R. (2009) 'Influenza A (H1N1) Virus', pp. 115–119.

Trovão, N. S. *et al.* (2015) 'Bayesian inference reveals host-specific contributions to the epidemic expansion of influenza A H5N1', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msv185.

Truong, T. C., Van Than, T. and Kim, W. (2014) 'Evolutionary phylodynamics of korean noroviruses reveals a novel gii.2/gii.10 recombination event', *PLoS ONE*, 9(12). doi: 10.1371/journal.pone.0113966.

Tumpey, T. M. *et al.* (2005) 'Characterization of the reconstructed 1918 Spanish influenza pandemic virus.', *Science (New York, N.Y.)*, 310(5745), pp. 77–80. doi: 10.1126/science.1119392.

Tuncer, N. and Le, T. (2014) 'Effect of air travel on the spread of an avian influenza pandemic to the United States', *International Journal of Critical Infrastructure Protection*, 7(1), pp. 27–47. doi: 10.1016/j.ijcip.2014.02.001.

Uchida, H. and Nelson, A. (2008) 'Agglomeration Index : Towards a New Measure of Urban', *World Development Report: Reshaping Economic Geography*, p. 19. Available at: http://siteresources.worldbank.org/INTWDR2009/Resources/4231006-1204741572978/Hiro1.pdf.

Ueda, M. *et al.* (2008) 'Maturation efficiency of viral glycoproteins in the ER impacts the

production of influenza A virus.', *Virus research*, 136(1–2), pp. 91–7. doi: 10.1016/j.virusres.2008.04.028.

Urbaniak, K. and Markowska-Daniel, I. (2014) 'In vivo reassortment of influenza viruses', *Acta Biochimica Polonica*, pp. 427–431.

Valeri, L. *et al.* (2016) 'Predicting subnational Ebola virus disease epidemic dynamics from sociodemographic indicators', *PLoS ONE*, 11(10). doi: 10.1371/journal.pone.0163544.

Valley-Omar, Z. *et al.* (2015) 'Phylogenetic Exploration of Nosocomial Transmission Chains of 2009 Influenza A/H1N1 among Children Admitted at Red Cross War Memorial Children's Hospital, Cape Town, South Africa in 2011.', *PloS one*, 10(11), p. e0141744. doi: 10.1371/journal.pone.0141744.

Vasin, A. V *et al.* (2016) 'The influenza A virus NS genome segment displays lineage - specific patterns in predicted RNA secondary structure', *BMC Research Notes*. BioMed Central, pp. 1–7. doi: 10.1186/s13104-016-2083-6.

Venter, M. *et al.* (2012) 'Evolutionary dynamics of 2009 pandemic influenza A virus subtype H1N1 in South Africa during 2009-2010.', *The Journal of infectious diseases*, 206 Suppl(Suppl 1), pp. S166-72. doi: 10.1093/infdis/jis539.

Venter, M. (2018) 'Assessing the zoonotic potential of arboviruses of African origin', *Current Opinion in Virology*. Elsevier B.V., 28(December 2017), pp. 74–84. doi: 10.1016/j.coviro.2017.11.004.

Viboud, C. *et al.* (2013) 'Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614), p. 20120199. doi: 10.1098/rstb.2012.0199.

Viboud, C., Alonso, W. J. and Simonsen, L. (2006) 'Influenza in Tropical Regions', *PLoS Medicine*, 3(4), p. e89. doi: 10.1371/journal.pmed.0030089.

Vijaykrishna, D. *et al.* (2010) 'Reassortment of pandemic H1N1/2009 influenza A virus in swine.', *Science (New York, N.Y.)*, 328(5985), p. 1529. doi: 10.1126/science.1189132.

Vijaykrishna, D., Holmes, Edward C, *et al.* (2015) 'The contrasting phylodynamics of human influenza B viruses', *eLife*, 4, pp. 1–23. doi: 10.7554/eLife.05055.

Vijaykrishna, D., Holmes, Edward C., *et al.* (2015) 'The contrasting phylodynamics of human influenza B viruses', *eLife*, 4, p. e05055. doi: 10.7554/eLife.05055.

Vijaykrishna, D., Mukerji, R. and Smith, G. J. D. (2015) 'RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion', *PLOS Pathogens*, 11(7), p. e1004902. doi: 10.1371/journal.ppat.1004902.

Villinger, J. *et al.* (2017) 'Arbovirus and insect-specific virus discovery in Kenya by novel six genera multiplex high-resolution melting analysis', *Molecular Ecology Resources*, 17(3), pp. 466–480. doi: 10.1111/1755-0998.12584.

'Virus-like-particles-as-a-vaccine-delivery-system-Myths-and-facts_2008_Human-Vaccines' (no date).

Vitoria, M. *et al.* (2009) 'The global fight against HIV/AIDS, tuberculosis, and malaria: Current status and future perspectives', *American Journal of Clinical Pathology*, 131(6), pp. 844–848. doi: 10.1309/AJCP5XHDB1PNAEYT.

Volz, E. M., Koelle, K. and Bedford, T. (2013) 'Viral Phylodynamics', *PLoS Computational Biology*, 9(3). doi: 10.1371/journal.pcbi.1002947.

Volz, E. M., Romero-Severson, E. and Leitner, T. (2017a) 'Phylodynamic Inference across Epidemic Scales', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msx077.

Volz, E. M., Romero-Severson, E. and Leitner, T. (2017b) 'Phylodynamic Inference across Epidemic Scales', *Molecular Biology and Evolution*, 34(5), pp. 1276–1288. doi: 10.1093/molbev/msx077.

Wan, Y. *et al.* (2011) 'Understanding the transcriptome through RNA structure', *Nature Reviews Genetics*. Nature Publishing Group, 12(9), pp. 641–655. doi: 10.1038/nrg3049.

Wang, X., Li, P. and Gutenkunst, R. N. (2017) 'Systematic Effects Of mRNA Secondary Structure On Gene Expression And Molecular Function In Budding Yeast', *bioRxiv*. Available at: http://biorxiv.org/content/early/2017/05/16/138792.abstract.

Washietl, S., Bernhart, S. H. and Kellis, M. (2014) 'Energy-based RNA consensus secondary structure prediction in multiple sequence alignments', *Methods in Molecular Biology*, 1097, pp. 125–141. doi: 10.1007/978-1-62703-709-9__7.

Watson, S. J. *et al.* (2015) 'Molecular Epidemiology and Evolution of Influenza Viruses Circulating within European Swine between 2009 and 2013', *Journal of Virology*, 89(July), p. JVI.00840-15. doi: 10.1128/JVI.00840-15.

Webster, R. *et al.* (1992) 'Evolution and ecology of influenza A viruses.', *Microbiological reviews*, 56(1), pp. 152–179. Available at: http://mmbr.asm.org/content/56/1/152.short.

Westgeest, K. B. *et al.* (2014) 'Genomewide analysis of reassortment and evolution of human influenza A(H3N2) viruses circulating between 1968 and 2011.', *Journal of virology*, 88(5), pp. 2844–57. doi: 10.1128/JVI.02163-13.

White, K. A., Enjuanes, L. and Berkhout, B. (2011) 'RNA virus replication, transcription and recombination.', *RNA biology*, 8(2), pp. 182–3. doi: 10.4161/rna.8.2.15663.

White, M. C., Steel, J. and Lowen, A. C. (2017) 'Heterologous Packaging Signals on Segment 4, but Not Segment 6 or Segment 8, Limit Influenza A Virus Reassortment', *Journal of Virology*, 91(11), pp. e00195-17. doi: 10.1128/JVI.00195-17.

WHO (2014) 'Meeting of the Strategic Advisory Group of Experts on Immunization, April 2014 - conclusions and recommendations', *Weekly epidemiological record*, 89(No. 21), pp. 221–236. doi: 10.1111/irv.12324/epdf.

Widdowson, M.-A., Iuliano, a D. and Dawood, F. S. (2014) 'Challenges to global pandemic mortality estimation.', *The Lancet. Infectious diseases*, 14(8), pp. 670–2. doi: 10.1016/S1473-3099(14)70835-7.

Wille, M. *et al.* (2013) 'Frequency and patterns of reassortment in natural influenza A virus infection in a reservoir host', *Virology*, 443(1), pp. 150–160. doi: 10.1016/j.virol.2013.05.004.

Wirth, T. *et al.* (2008) 'Origin, spread and demography of the Mycobacterium tuberculosis complex', *PLoS Pathogens*, 4(9). doi: 10.1371/journal.ppat.1000160.

Wong, K. K. *et al.* (2012) 'Epidemiology of 2009 pandemic influenza a virus subtype H1N1 among kenyans aged 2 months to 18 years, 2009-2010', *Journal of Infectious Diseases*,

206(SUPPL.1), pp. 2009–2010. doi: 10.1093/infdis/jis585.

Worby, C. J. *et al.* (2016) 'Reconstructing transmission trees for communicable diseases using densely sampled genetic data', *Annals of Applied Statistics*. doi: 10.1214/15-AOAS898.

Worobey, M. *et al.* (2008) 'Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960', *Nature*, 455(7213), pp. 661–664. doi: 10.1038/nature07390.

Worobey, M. *et al.* (2016) '1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America', *Nature*. doi: 10.1038/nature19827.

Worobey, M., Han, G.-Z. and Rambaut, A. (2014) 'Genesis and pathogenesis of the 1918 pandemic H1N1 influenza A virus.', *Proceedings of the National Academy of Sciences of the United States of America*, 111(22), pp. 8107–12. doi: 10.1073/pnas.1324197111.

Worobey, M., Han, G. and Rambaut, A. (2014) 'Genesis and pathogenesis of the 1918 pandemic H1N1 influenza A virus', 111(22). doi: 10.1073/pnas.1324197111.

Worobey, M. and Holmes, E. C. (1999) 'Evolutionary aspects of recombination in RNA viruses', *Journal of General Virology*, 80(10), pp. 2535–2543.

Wu, Q. *et al.* (2014) 'Identification of Viruses and Viroids by Next-Generation Sequencing and Homology-Dependent and Homology-Independent Algorithms', *Annual Review of Phytopathology*, 53(1), p. 150605182533006. doi: 10.1146/annurev-phyto-080614-120030.

Wu, X. *et al.* (2013) 'Impact of global change on transmission of human infectious diseases', *Science China Earth Sciences*, 57(2), pp. 189–203. doi: 10.1007/s11430-013-4635-0.

Xu, C. *et al.* (2014) 'Comparative Epidemiology of Influenza B Yamagata-and Victoria-Lineage Viruses in Households', *American Journal of Epidemiology*, 182(8), pp. 705–713. doi: 10.1093/aje/kwv110.

Xu, F., Connell McCluskey, C. and Cressman, R. (2013) 'Spatial spread of an epidemic through public transportation systems with a hub.', *Mathematical biosciences*, 246(1), pp. 164–75. doi: 10.1016/j.mbs.2013.08.014.

Xu, X. *et al.* (2004) 'Reassortment and evolution of current human influenza A and B viruses', *Virus Research*, 103(1–2), pp. 55–60. doi: 10.1016/j.virusres.2004.02.013.

Yamayoshi, S. *et al.* (2014) 'Virulence-Affecting Amino Acid Changes in the PA Protein of H7N9 Influenza A Viruses.', *Journal of virology*, 88(6), pp. 3127–34. doi: 10.1128/JVI.03155-13.

Yang, Z. and Rannala, B. (1997) 'Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method', *Mol. Biol. Evol.* , 14, pp. 717–724.

Yurovsky, A. and Moret, B. M. E. (2010) 'FluRF, an automated flu virus reassortment finder based on phylogenetic trees', *Proceedings - 2010 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2010*. BioMed Central Ltd, 12(Suppl 2), pp. 579–584. doi: 10.1109/BIBM.2010.5706632.

Zappa, A. *et al.* (2009) 'Emerging and re-emerging viruses in the era of globalisation', *Blood Transfusion*, 7(3), pp. 167–171. doi: 10.2450/2009.0076-08.

Zehender, G. *et al.* (2013) 'Reconstruction of the evolutionary dynamics of hepatitis C virus subtypes in Montenegro and the Balkan region', *Infection, Genetics and Evolution*, 17. doi:

10.1016/j.meegid.2013.04.003.

Zeldovich, K. B. *et al.* (2015) 'Positive Selection Drives Preferred Segment Combinations during Influenza Virus Reassortment', *Molecular Biology and Evolution*, 32(6), pp. 1519–1532. doi: 10.1093/molbev/msv044.

Zemora, G. *et al.* (2016) 'Segmented Structure of Separate and Transposable DNA and RNA Elements as Suggested by Their Size Distributions', *RNA Biol*. Taylor & Francis, 11(2), pp. 1–16. doi: 10.1038/srep38892.

Zhao, B. *et al.* (2015) 'Epidemiological study of influenza B in Shanghai during the 2009-2014 seasons: Implications for influenza vaccination strategy', *Clinical Microbiology and Infection*, 21(7), pp. 694–700. doi: 10.1016/j.cmi.2015.03.009.

Zheng, W. and Tao, Y. J. (2013) 'Structure and assembly of the influenza A virus ribonucleoprotein complex', *FEBS Letters*. Federation of European Biochemical Societies, 587(8), pp. 1206–1214. doi: 10.1016/j.febslet.2013.02.048.

Zhou, N. A. N. N. A. N. *et al.* (1999) 'Genetic Reassortment of Avian , Swine , and Human Influenza A Viruses in American Pigs', 73(10), pp. 8851–8856.

Zhu, W. *et al.* (2015) 'The repeated introduction of the H3N2 virus from human to swine during 1979–1993 in China', *Infection, Genetics and Evolution*, 33, pp. 20–24. doi: 10.1016/j.meegid.2015.04.001.

Zinszer, K. *et al.* (2017) 'Spatial Determinants of Ebola Virus Disease Risk for the West African Epidemic', *PLoS Currents*, pp. 1–14. doi: 10.1371/currents.outbreaks.b494f2c6a396c72ec24cb4142765bb95.

Zuker, M. (2003) 'Mfold web server for nucleic acid folding and hybridization prediction', *Nucleic Acids Research*, 31(13), pp. 3406–3415. doi: 10.1093/nar/gkg595.

APPENDICES

Appendix 1 Nucleotide sequence accession numbers of Influenza A/H1N1 pandemic 2009 virus HA, NA and MP sequences retrieved from GISAID database

HA

189113	217203	255225	256816	262420	262451	278364
189123	227681	255226	256817	262421	262452	278366
191939	227751	255227	262391	262422	262453	278368
191971	227845	255228	262392	262423	262454	278370
194180	227848	255229	262393	262424	262455	278372
194242	231308	255230	262394	262425	262456	278374
194251	232358	255231	262395	262426	262457	278376
210409	235545	255232	262396	262427	262458	278378
210442	235546	255233	262397	262428	269566	278380
210502	235550	255234	262398	262429	271976	278382
210504	235554	255235	262399	262430	271979	278384
210520	235556	255236	262400	262431	271982	278386
210522	235560	255237	262401	262432	271985	278388
210524	238325	255238	262402	262433	272031	278390
210526	238327	255239	262403	262434	272037	278392
210528	239637	255833	262404	262435	272040	278394
210530	239640	255835	262405	262436	272043	278396
210538	243775	255857	262406	262437	272046	278398
210544	243976	255859	262407	262438	272184	278400
210546	243978	255861	262408	262439	273851	278570
211651	243980	255987	262409	262440	273857	278576
211655	247266	255989	262410	262441	273908	278582
211663	247269	255991	262411	262442	278017	278584
211671	247456	256800	262412	262443	278025	278586
211681	252089	256802	262413	262444	278033	279899
215927	255219	256804	262414	262445	278352	279901
216798	255220	256806	262415	262446	278354	279903
217170	255221	256808	262416	262447	278356	279905
217173	255222	256810	262417	262448	278358	280347
217175	255223	256812	262418	262449	278360	280349
217178	255224	256814	262419	262450	278362	280352

280355	301704	317129	331040	355994	361631	361666
280358	301707	317132	331059	356256	361632	361667
283297	301904	317139	331061	356268	361633	361668
286988	301906	317142	331219	357499	361634	361669
286990	301924	317145	332598	357500	361635	361670
286992	301926	317148	335813	357501	361636	361671
286994	307143	317151	335816	357502	361637	361672
286996	307144	317154	335834	357503	361638	361673
286998	307145	317157	341948	357504	361639	361674
287000	307146	317160	341953	357505	361640	366281
287002	307147	317162	346645	357506	361641	375565
287004	307148	317165	346647	357507	361642	386007
287006	307149	317168	346649	357508	361643	386013
287008	307150	317171	346651	357509	361644	386019
287010	307151	317174	346653	357510	361645	386034
287012	307152	319430	346655	357511	361646	386036
287014	307153	319436	346683	357512	361647	386042
287016	307157	319437	346685	357513	361648	390479
287018	307158	319439	346687	358060	361649	390481
287020	307159	319441	346689	358064	361650	390483
295441	309639	319443	346691	358072	361651	397722
295444	309880	319471	346703	358074	361652	397725
295447	309922	319473	346705	358076	361653	397728
295504	309925	319592	346715	358078	361654	397731
295507	309928	319594	346717	358084	361655	397734
295510	309931	320033	346719	358086	361656	397737
295606	310058	325582	346721	361622	361657	405786
295608	316369	331002	346724	361623	361658	405796
295610	316442	331004	347045	361624	361659	405803
295612	316445	331006	347553	361625	361660	405853
295613	316932	331008	348177	361626	361661	405855
295615	316935	331010	352276	361627	361662	405856
300990	316938	331014	352293	361628	361663	416437
301413	317123	331036	352295	361629	361664	417122
301415	317126	331038	352303	361630	361665	417124

417126	439912	439947	443159	446142	457262	466630
417128	439913	439948	443169	446155	457263	466632
417130	439914	439949	443176	454436	457264	466634
418142	439915	439950	443184	457230	457265	477004
418145	439916	439951	443191	457231	457266	477006
422586	439917	439952	443198	457232	457267	477032
422613	439918	439953	443205	457233	457268	477034
422616	439919	439954	443212	457234	457269	482162
422627	439920	439955	443219	457235	457270	482169
422630	439921	439956	443226	457236	457271	484718
422633	439922	439957	443233	457237	457272	484954
422636	439923	439958	443241	457238	457273	487664
422639	439924	439959	443248	457239	457274	487666
422642	439925	439960	443255	457240	457275	487668
425962	439926	439961	443263	457241	460646	487670
426441	439927	439962	443271	457242	466551	487672
426450	439928	439963	443278	457243	466564	487674
432618	439929	439964	443285	457244	466566	487676
432620	439930	439965	446012	457245	466568	487678
438813	439931	439966	446020	457246	466570	487680
438815	439932	439967	446029	457247	466572	491208
439241	439933	439968	446036	457248	466574	491211
439242	439934	439969	446043	457249	466576	491214
439247	439935	439970	446050	457250	466578	491220
439460	439936	439971	446057	457251	466580	498409
439902	439937	439972	446064	457252	466582	498411
439903	439938	439973	446071	457253	466612	498417
439904	439939	439974	446078	457254	466614	498419
439905	439940	439975	446086	457255	466616	498427
439906	439941	439976	446094	457256	466618	498429
439907	439942	440003	446102	457257	466620	498431
439908	439943	440929	446110	457258	466622	499320
439909	439944	443017	446118	457259	466624	503849
439910	439945	443067	446126	457260	466626	503852
439911	439946	443120	446134	457261	466628	503855

504755	507829	516945	531928	536865	539482	541029
504758	507831	516965	533526	538146	539484	541031
504781	513253	516967	536247	538152	539486	541033
504788	515423	516969	536249	539470	539532	541035
504793	515425	516971	536251	539472	539534	541037
505297	515723	516979	536253	539474	539536	
507551	516941	516981	536255	539476	541025	
507827	516943	516983	536257	539478	541027	

NA
189114
189124
191940
194181
194243
194252
210410
210503
210505
210521
210523
210525
210527
210529
210531
210539
210545
210547
211650
211654
211662
211670
211680
216797
217169
217172
217177
217202
227680
227750
227844
227847
231307
231342
232357
235549
235553
235555
235559
238324
238326
239636
239639
243975
243977
243979
247265
247268
247453
247455
255832
255834
255856
255858
255860
255986
255988
255990
256799
256801
256803
256805
256807
256809
256811
256813
256815
271975
271978
271981
271984
272030
272036
272039
272042
272045
272299
273850
273856
273907
278019
278027
278035
278351
278353
278355
278357
278359
278361
278363
278365
278367
278369
278371
278373
278375
278377
278379
278381
278383
278385
278387
278389
278391
278393
278395
278397
278399
278569
278575
278581
278583
278585
279898
279900
279902
279904
280346
280351
280354
280357
284114
284115
284116
284117
284118
284119
284120
284121
284122
284123
284124
284125
284126
284127
284128
284129
284130
284131
284132
284133
284134
284135
284136
284137
284138
284139
284140
284141
284142
284143
284144
284145
284146
284147
284148
284149
284150
284151
284152
284153
284154
284155
284156
284157
284158
284159
286987
286989
286991
286993
286995
286997
286999
287001
287003
287005
287007
287009
287011
287013
287015
287017
287019
295440
295443
295446
295503
295506
295509
295607
295609
295611
295614
295616
300989
301412
301414
301703
301706
301903
301905
301923
301925
309640
309879
309921
309924
309927
309930
310057
316368
316441
316444
316933
316936
316939
317124
317127
317130
317133
317140
317143
317146
317149
317152
317155
317158
317161
317163
317166
317169
317172
317175
319431
319438
319440
319442
319444
319472
319474
319593
319595
320032
325581
331003
331005
331007
331009
331011
331015
331037
331039
331041
331060
331062
331218
332597
335812
335815
335833
341612
341952
346646
346648
346650
346652
346654
346656
346684
346686
346688
346690
346692
346704
346706
346716
346718
346720
346725
347044
347552
348176
352294
352296
352304
355005
356255
356267
357514
358061
358065
358073
358075
358077
358079
358085
358087
366280
375564
386012
386033
386035
386041
390480
390482
390484
392897
392904

397721	443171	461524	487677	536252
397724	443178	461525	487679	536254
397727	443186	461526	487681	536256
397730	443193	461527	491207	536258
397733	443200	461528	491210	536864
397736	443207	461529	491213	538145
405787	443214	461530	491219	538151
405797	443221	461531	498410	539471
405804	443228	466552	498412	539473
405854	443235	466565	498418	539475
405857	443243	466567	498420	539477
416438	443250	466569	498428	539479
417123	443257	466571	498430	539483
417125	443265	466573	498432	539485
417127	443273	466575	498873	539487
417129	443280	466577	503848	539533
417131	443287	466579	503851	539535
418141	446014	466581	503854	539537
418144	446022	466583	504754	541026
422585	446031	466613	504757	541028
422612	446038	466615	504780	541030
422615	446045	466617	504787	541032
422626	446052	466619	504792	541034
422629	446059	466621	505298	541036
422632	446066	466623	507828	541038
422635	446073	466625	507830	
422638	446080	466627	507832	
422641	446088	466629	513252	
425963	446096	466631	515424	
426440	446104	466633	515426	
426449	446112	466635	515722	
426452	446120	477005	516942	
432617	446128	477007	516944	
432619	446136	477033	516946	
438814	446144	477035	516966	
438816	446157	477427	516968	
439240	454435	482161	516970	
439459	460647	482168	516972	
439998	461517	484953	516980	
439999	461518	487665	516982	
440002	461519	487667	516984	
443019	461520	487669	531927	
443069	461521	487671	533525	
443141	461522	487673	536248	
443161	461523	487675	536250	

MP
189117
189126
191943
194247
194256
216793
217168
217171
217174
217176
217198
227679
227749
227843
227846
231306
231309
239635
239638
240358
247264
247267
247452
247454
271974
271977
271980
271983
272029
272035
272038
272041
272044
273849
273855
273906
278020
278028
278036
280345
280348
280350
280353
280356
284160
284161
284162
284163
284164
284165
284166
284167
284168
284169
284170
284171
284172
284173
284174
284175
284176
284177
284178
284179
284180
284181
284182
284183
284184
284185
284186
284187
284188
284189
284190
284191
284192
284193
284194
284195
284196
284197
284198
284199
284200
284201
284202
284203
284204
284205
295439
295442
295445
295502
295505
295508
300988
301702
301705
309878
309923
309926
309929
310056
316367
316440
316443
316934
316937
316940
317125
317128
317131
317135
317141
317144
317147
317150
317153
317156
317159
317164
317167
317170
317173
317176
317177
320031
325564
325580
331217
332596
335811
335814
335832
341947
341951
347551
348175
355004
355653
355654
355658
356254
356266
357518
366279
376386
386011
387940
387944
387945
387948
387951
397720
397723
397726
397729
397732
397735
418140
418143
422584
422611
422614
422625
422628
422631
422634
422637
422640
426439
426448
426451
439246
439252
439519
439996
439997
443020
443070
443144
443163
443172
443179
443187
443194
443201
443208
443215
443222
443229
443236
443244
443251
443258
443266
443274
443281
443288
446015
446023
446032
446039
446046
446053
446060
446067
446074
446081
446089
446097
446105
446113
446121
446129
446137
446145
446158
454434
466661
466665
466667
466671
466672
466673
466674
466675
466678
466679
466680
466681
466682
482160
482167
484717
484952
491206
491209
491212
491218
498872
500449
500464
500470
500475
500490
500500
500530
503847
503850
503853
504753
504756
504779
504786
507909
507910
508200
508201
508202
508206
508207
508208
508209
508210
508211
508212
508213
508214
508215
508216
508217
508218
508219
508286
508287
508290
508291
508337
508338
508490
508492
508496
508497
508498
508499
508500
508501
509487
513251
515721
531926
533524
536863
538144
538150
539948
539988
539989

Appendix 2: Database search queries used to retrieve Influenza A, B and C virus genome set sequences used for reassortment and secondary structure analyses

Influenza A virus sequences
https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?cmd=show_query&country=any&fyear=1927&genomeset=full&go=genomeset&host=any&lab=exclude&lineage=include&niaid=include&qgenomeset=full&searchin=strain&sequence=N&subtype_h=any&subtype_mix=include&subtype_n=any&swine=include&tyear=2013&type=a&vac_strain=include

Influenza B Virus

https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?cmd=show_query&country=any&fpyear=1927&genomeset=full&go=genomeset&host=any&lab=exclude&lineage=include&niaid=include&qgenomeset=full&searchin=strain&sequence=N&subtype_h=any&subtype_mix=include&subtype_n=any&swine=include&tpyear=2013&type=b&vac_strain=include

Influenza C virus
https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?cmd=show_query&country=any&fpyear=1927&genomeset=full&go=genomeset&host=any&lab=exclude&lineage=include&niaid=include&qgenomeset=full&searchin=strain&sequence=N&subtype_h=any&subtype_mix=include&subtype_n=any&swine=include&tpyear=2013&type=c&vac_strain=include

Appendix 3 Influenza C virus isolates whose genome sets were retrieved from GISAID database

EPI_ISL_167197
EPI_ISL_176788
EPI_ISL_176787
EPI_ISL_176785
EPI_ISL_176784
EPI_ISL_176783
EPI_ISL_176782
EPI_ISL_176781
EPI_ISL_176780
EPI_ISL_176779
EPI_ISL_176778
EPI_ISL_176777
EPI_ISL_176776
EPI_ISL_176775
EPI_ISL_176774
EPI_ISL_176773
EPI_ISL_176772
EPI_ISL_144547
EPI_ISL_79749
EPI_ISL_79748
EPI_ISL_79747
EPI_ISL_79746
EPI_ISL_66443
EPI_ISL_66441
EPI_ISL_66439
EPI_ISL_66438
EPI_ISL_66432
EPI_ISL_66430
EPI_ISL_66429
EPI_ISL_66428
EPI_ISL_66427
EPI_ISL_66426
EPI_ISL_66425
EPI_ISL_66422
EPI_ISL_66420
EPI_ISL_66419
EPI_ISL_66418
EPI_ISL_66417
EPI_ISL_66416
EPI_ISL_66415
EPI_ISL_66413
EPI_ISL_66412
EPI_ISL_66411
EPI_ISL_66410
EPI_ISL_66408
EPI_ISL_66407
EPI_ISL_66406
EPI_ISL_66405
EPI_ISL_66404
EPI_ISL_66403
EPI_ISL_66401
EPI_ISL_66400
EPI_ISL_66399
EPI_ISL_66397
EPI_ISL_66396
EPI_ISL_66395
EPI_ISL_66394
EPI_ISL_66392
EPI_ISL_66391
EPI_ISL_66389
EPI_ISL_66388
EPI_ISL_66387
EPI_ISL_66386
EPI_ISL_66385
EPI_ISL_66384
EPI_ISL_66383
EPI_ISL_66382
EPI_ISL_66381
EPI_ISL_66380
EPI_ISL_66379
EPI_ISL_66378
EPI_ISL_66376
EPI_ISL_66375
EPI_ISL_66373
EPI_ISL_66371
EPI_ISL_66369
EPI_ISL_66363
EPI_ISL_66362
EPI_ISL_66361
EPI_ISL_66360
EPI_ISL_66358
EPI_ISL_66357
EPI_ISL_66356
EPI_ISL_66355
EPI_ISL_66353
EPI_ISL_66351
EPI_ISL_66349
EPI_ISL_66347
EPI_ISL_66344
EPI_ISL_66337
EPI_ISL_66336
EPI_ISL_66334
EPI_ISL_66331
EPI_ISL_66329
EPI_ISL_66328
EPI_ISL_66326
EPI_ISL_66325

Appendix 4: Nucleotide sequence accession numbers of Infectious Salmon Anemia virus used in this analysis

AF220607	AY853945	EU118818	JN710871	JN710927
AF302799	AY853946	EU118819	JN710872	JN710928
AF302801	AY853947	EU118820	JN710873	JN710929
AF302802	AY853948	EU851044	JN710874	JN710930
AF302803	AY853949	FJ178189	JN710875	JN710931
AF364869	AY853951	FN687348	JN710876	JN710932
AF364872	AY853952	FN687351	JN710877	JN710933
AF364873	AY853953	FN687352	JN710878	JN710934
AF364874	AY853954	FN687356	JN710879	JN710935
AF364879	AY853955	GU830897	JN710880	JN710936
AF364880	AY853956	GU830898	JN710881	JN710937
AF364881	AY853958	GU830899	JN710882	JN710938
AF364883	AY853962	GU830900	JN710883	JN710939
AF364885	AY853963	GU830905	JN710884	JN710940
AF364886	AY853965	GU830906	JN710885	JN710941
AF364887	AY853967	GU830907	JN710886	JN710942
AF364888	AY853968	GU830908	JN710887	JN710943
AF364889	AY853969	HQ664992	JN710888	JN710944
AF364890	AY971659	HQ664993	JN710889	JN710945
AF364891	AY971662	HQ664995	JN710890	JN710946
AF364892	AY971663	HQ664997	JN710891	JN710947
AF364893	AY971664	JN710835	JN710892	JN710948
AF364894	AY971666	JN710836	JN710893	JN710949
AF364895	AY971667	JN710837	JN710894	JN710950
AF364896	AY973179	JN710838	JN710895	JN710951
AF364897	AY973184	JN710839	JN710896	JN710952
AF526263	AY973190	JN710840	JN710897	JN710953
AY127875	AY973194	JN710841	JN710898	JN710954
AY127876	DQ108598	JN710842	JN710899	JN710955
AY127877	DQ108599	JN710843	JN710900	JN710956
AY127881	DQ108600	JN710844	JN710901	JN710957
AY601904	DQ108604	JN710845	JN710902	JN710958
AY744390	DQ108605	JN710846	JN710903	JN710959
AY744391	DQ108606	JN710847	JN710904	JN710960
AY744392	DQ108607	JN710848	JN710905	JN710961
AY853917	DQ785205	JN710849	JN710906	JN710962
AY853919	DQ785206	JN710850	JN710907	JN710963
AY853920	DQ785210	JN710851	JN710908	JN710964
AY853921	DQ785214	JN710852	JN710909	JN710965
AY853922	DQ785216	JN710853	JN710910	JN710966
AY853923	DQ785219	JN710854	JN710911	JN710967
AY853924	DQ785220	JN710855	JN710912	JN710968
AY853925	DQ785224	JN710856	JN710913	JN710969
AY853926	DQ785228	JN710857	JN710914	JN710970
AY853928	DQ785230	JN710859	JN710915	JN710971
AY853929	DQ785233	JN710860	JN710916	JN710972
AY853930	DQ785234	JN710861	JN710917	JN710973
AY853932	DQ785238	JN710862	JN710918	JN710974
AY853933	DQ785242	JN710863	JN710919	JN710975
AY853934	DQ785244	JN710864	JN710920	JN710976
AY853935	DQ785247	JN710865	JN710921	JN710977
AY853936	DQ785248	JN710866	JN710922	JN710978
AY853939	DQ785252	JN710867	JN710923	JN710979
AY853942	DQ785256	JN710868	JN710924	JN710980
AY853943	DQ785258	JN710869	JN710925	JN710981
AY853944	EU118817	JN710870	JN710926	JN710982

JN710983	JN711006	JN711029	JN711052	JN711075
JN710984	JN711007	JN711030	JN711053	JN711076
JN710985	JN711008	JN711031	JN711054	JN711077
JN710986	JN711009	JN711032	JN711055	JN711078
JN710987	JN711010	JN711033	JN711056	JN711079
JN710988	JN711011	JN711034	JN711057	JN711080
JN710989	JN711012	JN711035	JN711058	JN711081
JN710990	JN711013	JN711036	JN711059	JN711082
JN710991	JN711014	JN711037	JN711060	JN711083
JN710992	JN711015	JN711038	JN711061	JN711084
JN710993	JN711016	JN711039	JN711062	JN711085
JN710994	JN711017	JN711040	JN711063	JN711086
JN710995	JN711018	JN711041	JN711064	JN711087
JN710996	JN711019	JN711042	JN711065	JN711088
JN710997	JN711020	JN711043	JN711066	JN711089
JN710998	JN711021	JN711044	JN711067	JN711090
JN710999	JN711022	JN711045	JN711068	JN711091
JN711000	JN711023	JN711046	JN711069	JN711092
JN711001	JN711024	JN711047	JN711070	JN711093
JN711002	JN711025	JN711048	JN711071	JN711094
JN711003	JN711026	JN711049	JN711072	JN711095
JN711004	JN711027	JN711050	JN711073	JN711096
JN711005	JN711028	JN711051	JN711074	

Appendix 5: Nucleotide sequence accession numbers of Thogoto virus used in this analysis

M77280
AF236794
NC 6508
NC6507
NC6506
NC6504
NC6496
NC6495
AF006073
AF004985
AF527531
AF527530
AF527529
AF168988
AF168987
AF168986
AF168985
AF168984
AF168983
AF168982
AF168981
AF168980
AF168979
AF168978
AF168977
AF168976
AF168975
AF168974
AF168973
AF168972
AF168971
AF168970
AF168969
AF168968
AF168967
AF168966
AF168965
AF168964
AF168963
AF168962
AF168961
AF168960
AF168959
AF168958
AF168957
Y17873
X96872
D00540

Appendix 6: Tables illustrating Reassortment events detected within concatenated genomes of Influenza A, B and C viruses

	In Alignme		segments	Event foun	Detection Methods		
Event	Begin	End			RDP	GENECON	Maxchi
1	1*	6836	PB2, PB1,	35	4.2857708	1.9642942	1.9487858
2	6836	8602	HA	15	4.7350739	8.9597094	3.9940855
3	10143*	11614	NA	6	3.5616742	2.2883526	NS
4	10143*	11614	NA	86	9.7759231	2.3562651	NS
5	10143	11614*	NA	5	2.1997424	1.0094474	NS
6	4633	10143*	PA, HA, N	7	2.7112875	1.2439428	5.5716490
7	10143	11614*	NA	6	4.2889877	6.9982086	NS
8	6836	8602*	HA	21	1.2210429	2.3268534	NS
9	6836*	8602	HA	10	5.3123964	2.2629870	NS
10	9720	11614*	HA	11	5.8035177	3.3634900	5.5486897
11	6850*	8595*	HA	16	4.8196948	6.1347659	NS
12	10143	11614*	NA	3	7.3194428	3.2023226	6.5509238
13	6798*	11614	HA, NP, N	1	2.4453665	4.1526223	NS
14	6836*	8602	HA	25	3.7856390	8.6426301	NS
15	10143	11614	NA	13	2.0671661	1.0383210	NS
16	10143*	11614	NA	5	5.9088732	1.4918161	NS
17	10143*	11614	NA	4	2.4677066	7.7570373	4.0759678
18	10143	11614*	NA	9	1.3256362	2.1154168	4.5185502
19	10143	11614*	NA	35	9.9696850	1.1894674	NS
20	6836*	8602*	HA	27	3.4789059	6.7211041	NS
21	6836	8602*	HA	19	9.4207009	1.1248401	NS
22	6836*	8602*	HA	24	7.3512831	8.2528593	1.0584317
23	8602*	10143	NP	1	1.8097692	1.6813043	NS
24	12606	13466*	NS	19	1.9450466	2.8551036	NS
25	6836*	8602*	HA	5	5.0546819	9.0379792	3.5846200
26	6836*	8602*	HA	2	7.9040390	NS	NS
27	12606*	13466*	NS	9	1.0322929	1.8363140	NS
28	6836*	8602*	HA	1274	4.7731105	3.4677235	3.7875568
29	1*	6836	PB2, PB1,	203	1.8489916	1.5360995	NS
30	12606*	13466*	NS	33	7.9765419	1.3924662	NS
31	2319*	4633*	PB1	1	4.4096753	7.1615799	1.7892083
32	10143*	11614*	NA	211	5.0366816	1.7352194	2.7851593
33	6836	8602*	HA	1	3.5119901	5.1217023	NS
34	10143*	11614*	NA	4	2.1282677	3.6827402	4.5436625
35	4633*	6836*	PA	3	1.1696346	4.8999267	NS
36	6836*	8602*	HA	3	1.6511939	4.0325976	NS
37	10143*	11614*	NA	6	6.6078701	5.4527441	NS
38	10143*	11614*	NA	562	6.2747317	1.8687928	3.0116971
39	10143*	11614*	NA	866	4.2893693	1.9015221	4.7943272
40	10143	11614	NA	7	6.3105912	2.0358378	NS
41	12606*	13466*	NS	53	6.4932370	1.7845552	9.9050514
42	6836*	8602*	HA	17	2.4760200	1.0941661	1.8468609
43	6836*	8602*	HA	2	1.6784624	9.4400080	NS
44	6836*	8602*	HA	4	2.7614129	6.9103548	NS
45	6836	8602*	HA	9	1.0969536	1.5871964	NS
46	6836*	8602*	HA	34	3.5730873	1.0882374	1.1394138
47	2319*	4633*	PB1	596	1.5031653	4.4292039	NS
48	12606*	13466*	NS	49	1.4365557	3.2337503	NS
49	2319*	4633*	PB1	1	4.0771393	NS	NS
50	10164*	11590*	NA	443	8.0877104	1.3312626	1.3790288
51	10143	11614*	NA	1	4.2509908	5.4960432	NS
52	4633*	6836*	PA	11	2.1944624	1.1299325	8.2490000
53	6836*	8602*	NA	1	NS	1.1146624	1.6828509
54	6836*	8602*	NA	22	1.6822668	1.0352717	4.2844195
55	12606*	13466*	NS	12	2.7344893	2.9836583	NS
56	10143*	11614*	NA	46	1.5227870	5.4355794	NS
57	12606*	13466*	NS	1	6.6224218	1.6351888	1.0278061
58	10143*	11614*	NA	1	4.8112307	1.8666537	2.9212522
59	10143	11614*	NA	9	2.7834981	4.1427136	7.0286067
60	6865*	7356	HA	1	7.5440169	2.4471234	1.7992117
61	12606*	13466*	NS	3	3.8003854	5.9344858	NS
62	89*	2193*	PB2	8	6.8945888	1.2737439	NS
63	10143*	11614*	NA	5	1.3644254	5.0009635	1.3729646
64	6836*	8602*	HA	5	1.2115627	5.9856336	5.1511042
65	2319*	6836*	PB1, PA	1522	NS	7.4701353	1.1553911
66	1*	2319*	PB2	1	1.0420408	2.1526654	1.5135189
67	8602*	10143*	NP	1	5.9330160	5.2908584	NS
68	6836*	8602*	HA	2	2.5234957	2.4979109	4.9178555
69	1*	2319*	PB2	6	3.0264265	2.9379047	7.6683953
70	8602*	10143*	NP	2	2.0937301	9.1148018	NS
71	8602*	10143*	NP	5	6.1365873	5.5516779	NS
72	8602*	10143*	NP	34	3.1327977	1.5900914	NS
73	1*	2319*	PB2	4	1.3562555	2.2067545	4.5030974
74	11614*	12606*	MP	5	9.1680893	2.2572314	NS
75	8590*	10142*	NP	1	1.3699885	4.3478139	8.0189027
76	12606*	13466*	NS	14	6.2956018	5.5065472	4.7120333
77	8607*	13470*	NP, NA, M	72	1.6939858	1.9363385	NS
78	8602*	10143*	NP	1	8.3277915	1.1300675	NS
79	12606*	13466*	NS	1	2.9533994	1.4708005	1.1295529
80	4633*	6836*	PA	8	1.0492511	8.5590450	7.9017368
81	1*	2319*	PB2	1	6.5773871	2.3280732	NS
82	6836*	11614*	HA, NP, N	1	1.1111657	5.6512720	5.2492918
83	8602*	10143*	HA, NP	189	4.0095847	1.6449434	NS
84	1*	2319*	PB2	104	1.8194779	3.8019066	2.6667375

	Break point In Alignment			Detection Methods			
Event	Begin	End	Segments involved	Detected in seq(s)	RDP	GENECON	Maxchi
1	6814	8844	HE	34	0.0082700	NS	4.91E-13
2	6814	8844	HE	1	3.63E-11	3.17E-08	1.91E-05
3	11776	12704*	NS1/NS2	1	8.25E-11	1.08E-09	7.24E-05
4	6814	10601	HE, NP	1	3.99E-09	1.75E-06	1.40E-07
5	11776	12704*	NS1/NS2	1	9.91E-09	2.35E-08	0.0006660
6	2325	4659	PB1	1	0.0011502	1.92E-07	NS
7	7432	7868*	HE	1	0.0006196	1.50E-05	0.0388803
8	7324*	8577*	HE	1	4.33E-05	NS	1.66E-05

Event	Begin	End	Segments involved	Detected in seq(s)	RDP	GENECONV	Maxchi
1	1*	2344	PB2	1	5.8180731	3.51908125638	1.7371062
2	8794*	10597*	NP	1	9.8085575	3.85008673862	NS
3	6954	8794	HA	16	1.7638548	2.37861664312	5.1622015
4	2344	4703*	PB1	13	2.3992485	3.43654172206	NS
5	3338	4703	PB1	1	4.1986240	1.37008038088	1.8209693
6	8724	10597	NP	333	1.6727993	1.24399290073	NS
7	10597	14316*	NA, M1/M2, NS1/NS2	20	4.2038043	1.26876192422	1.8232293
8	4703*	6954*	PA	8	7.3574973	4.04041582641	1.8704556
9	10597	14316*	NA, M1/M2, NS1/NS2	5	5.3332712	8.67060427006	3.0413989
10	8766*	10597*	NP	292	5.8981568	6.53752767726	1.2022897
11	4703*	6954*	PA	3	5.4215131	5.85154863143	1.0128543
12	10597*	12123*	NA	18	1.7659181	NS	NS
13	13276*	14316*	NS1/NS2	22	1.2569503	3.22052909145	NS
14	3440	3794*	PB1	1	1.0032787	2.17508306247	.01511003
15	10597*	12123*	NA	11	3.8062324	6.63403828240	5.9727326
16	1*	2344*	PB2	569	4.9230610	1.10201131514	5.6500065
17	12123*	13276*	M1/M2	54	1.0162586	3.27914089210	NS
18	8719	10597*	NP	6	8.9667169	6.38740113778	NS
19	6954*	8794*	HA	1	1.8919105	NS	1.3469569
20	13276*	14316*	NS1/NS2	2	4.1068356	NS	9.4426212
21	8794*	10597*	NP	4	7.6413822	1.41344189103	NS
22	4703*	6954*	PA	547	1.0537816	8.58694675737	NS
23	2344*	4703*	PB1	11	2.6231288	7.98671501455	7.7541614
24	10597*	14316*	NA, M1/M2, NS1/NS2	2	.00547953	NS	1.3190018
25	13276*	14316*	NS1/NS2	728	3.1658399	2.91477255637	NS
26	12123	13276*	M1/M2	1	4.7673421	1.77521795165	NS
27	10597*	12123*	NA	5	5.8465334	5.51261546932	NS
28	1*	2344*	PB2	82	NS	NS	4.4750938
29	10597	13276*	NA, M1/M2	62	NS	.009043831704	5.1415233
30	4012	8794*	PB1, PA, HA	440	4.6498721	1.11766111925	1.1382798
31	2405*	4006	PB1	126	2.1609668	1.40544820857	3.3563488
32	1*	2344*	PB2	4	.01518141	NS	1.5302881
33	4703	6954*	PA	12	.00666261	NS	1.4398676
34	1*	4000	PB2, PB1	5	.00160604	NS	1.6999177
35	2344*	4703*	PB1	2	5.9248099	2.09367435785	NS
36	2344*	4703*	PB1	1	2.2792474	.002649929046	NS
37	10597*	12123*	NP	40	7.8839529	.013261936301	NS
38	2344*	4703*	PB1	1	.00015087	NS	NS
39	10597*	12123*	NP	44	.00016134	.011165495791	NS
40	6954*	13276*	HA, NP, NA, M1/M2	47	NS	NS	.00078783

Appendix 7 Maps of sites detected to be folded into secondary structures along Individual homologous segments in Influenza A, ,B C, ISAV and Thogoto Viruses

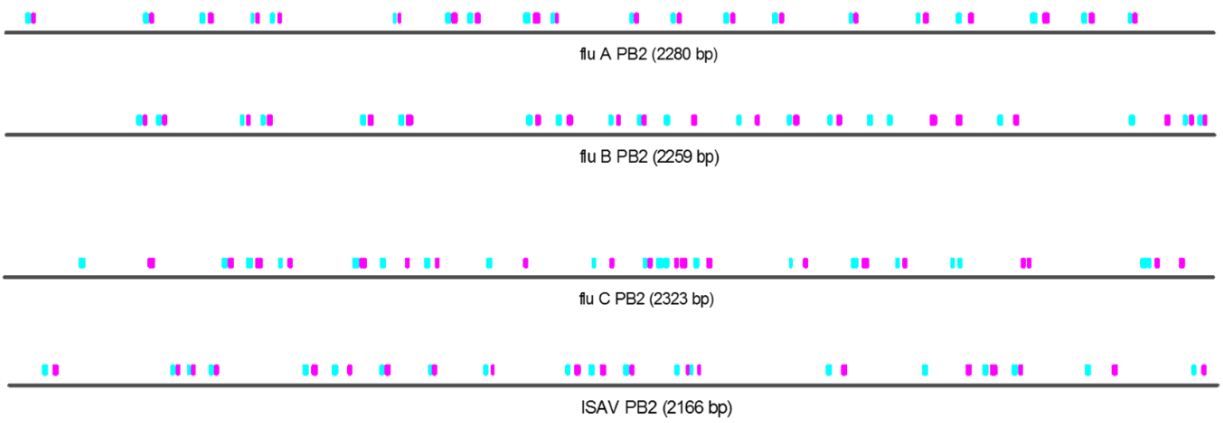


Figure 4.1 Predicted RNA secondary structure map along segments 1 (Influenza A-PB2, Influenza B-PB2, Influenza C-PB2 and Isavirus-PB2). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs. From the maps, it is evident some regions of the genomes are predicted to be structured than others and this may impact on their probable molecular stability.

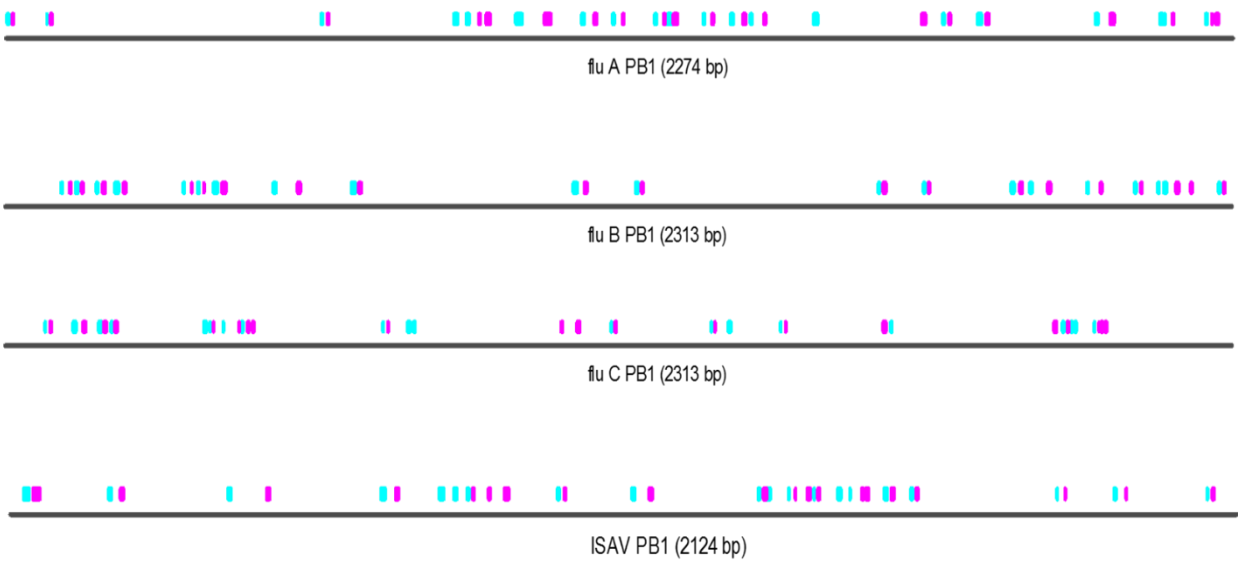


Figure 4.2 Predicted RNA secondary structure map along segments 2 (Influenza A-PB1, Influenza B-PB1, Influenza C-PB1 and Isavirus-PB1). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs.

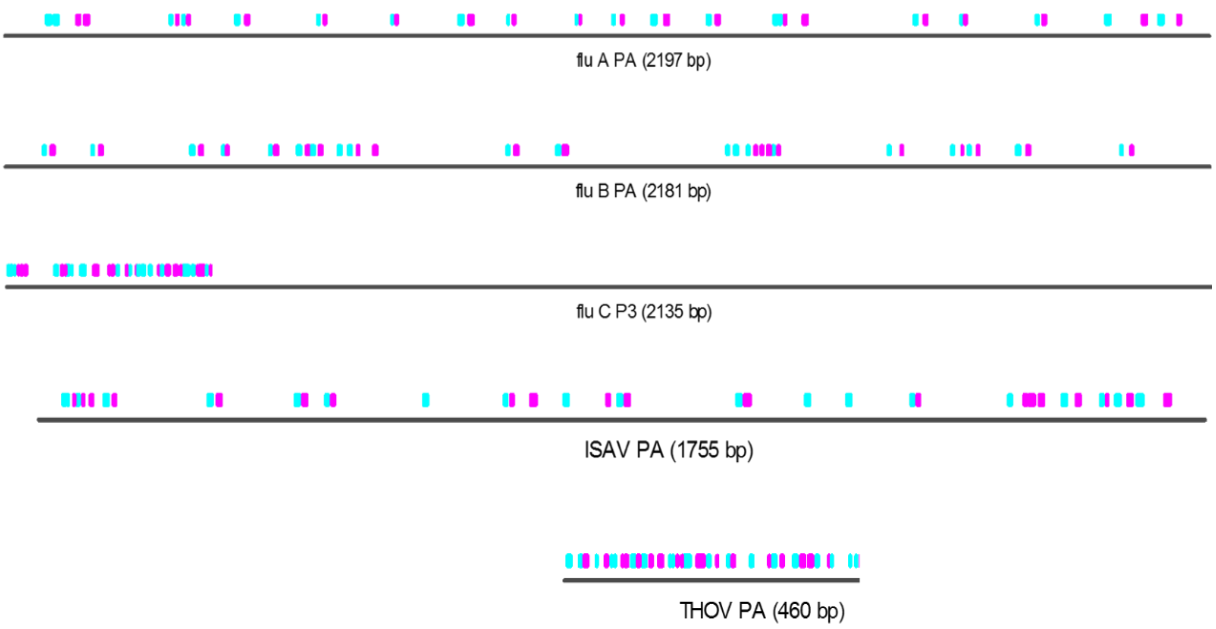


Figure 4.3 Predicted RNA secondary structure map along orthomyxovirus segments 3 (Influenza A-PA, Influenza B-PA, Influenza C-P3, Isavirus-PA and THOV-PA). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs.

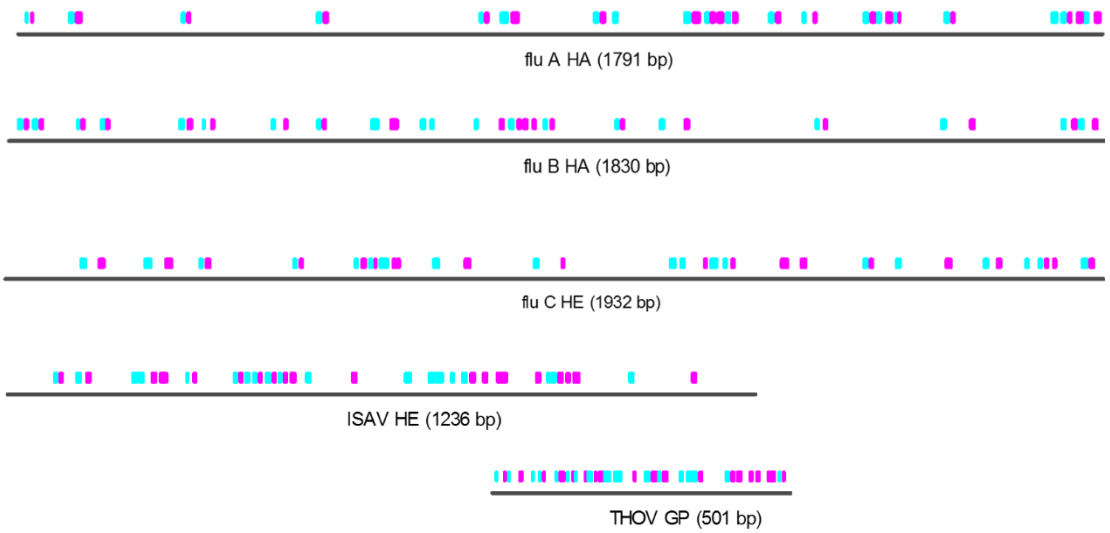


Figure 4.4 Predicted RNA secondary structure map along orthomyxovirus segment 4 (Influenza A-HA, Influenza B-HA, Influenza C-HE, Isavirus-HE and THOV-GP). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs.

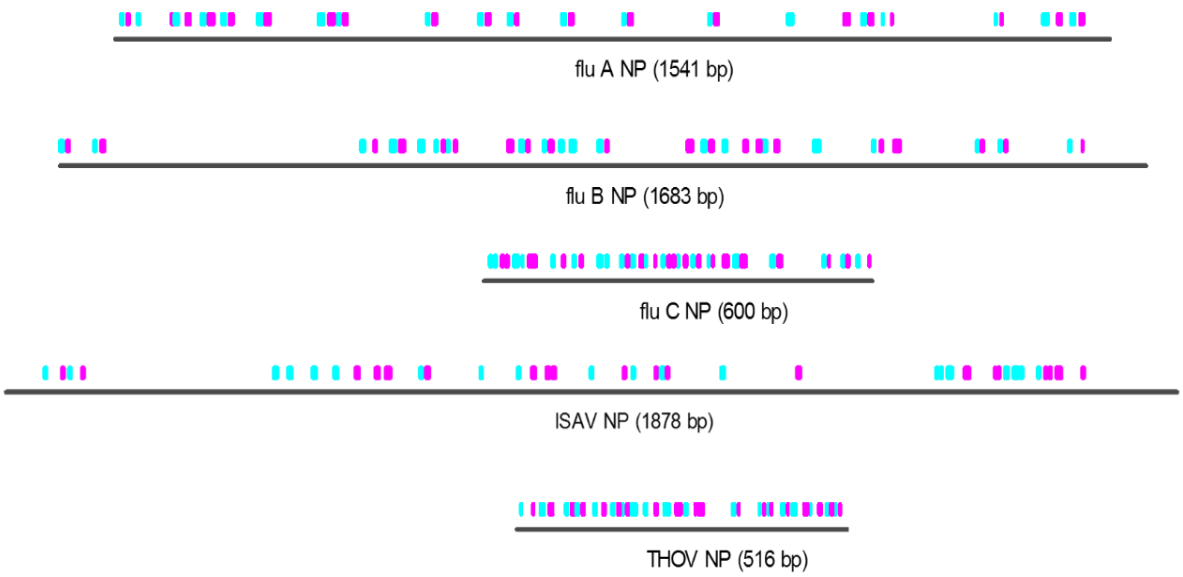


Figure 4.5 Predicted RNA secondary structure map along orthomyxovirus segment 5 (Influenza A-NP, Influenza B-NP, Influenza C-NP, Isavirus-NP and THOV-NP). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs.

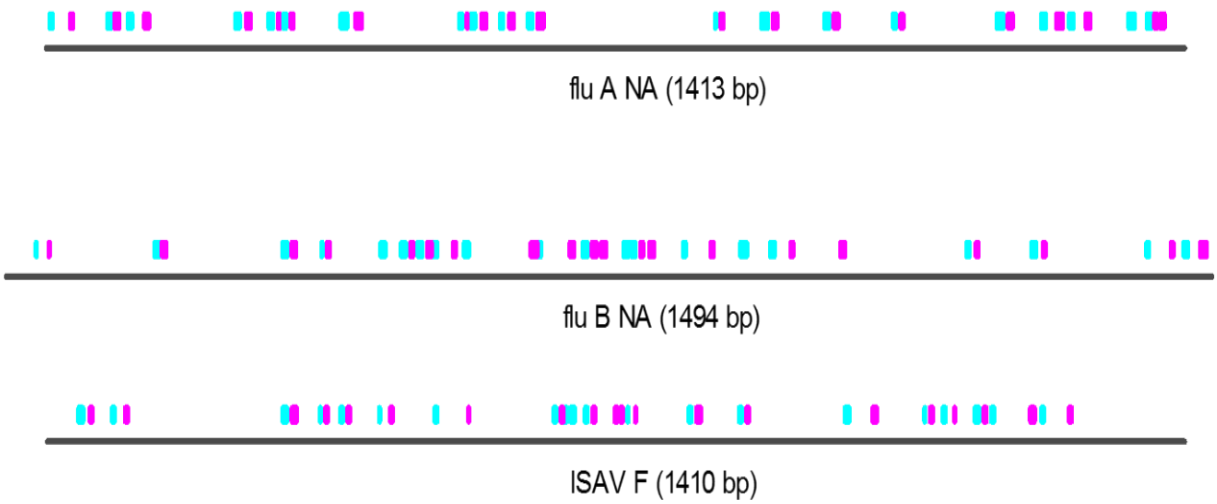


Figure 4.6 Predicted RNA secondary structure map along orthomyxovirus segment 6 (Influenza A-NA, Influenza B-NA, and Isavirus-F). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs

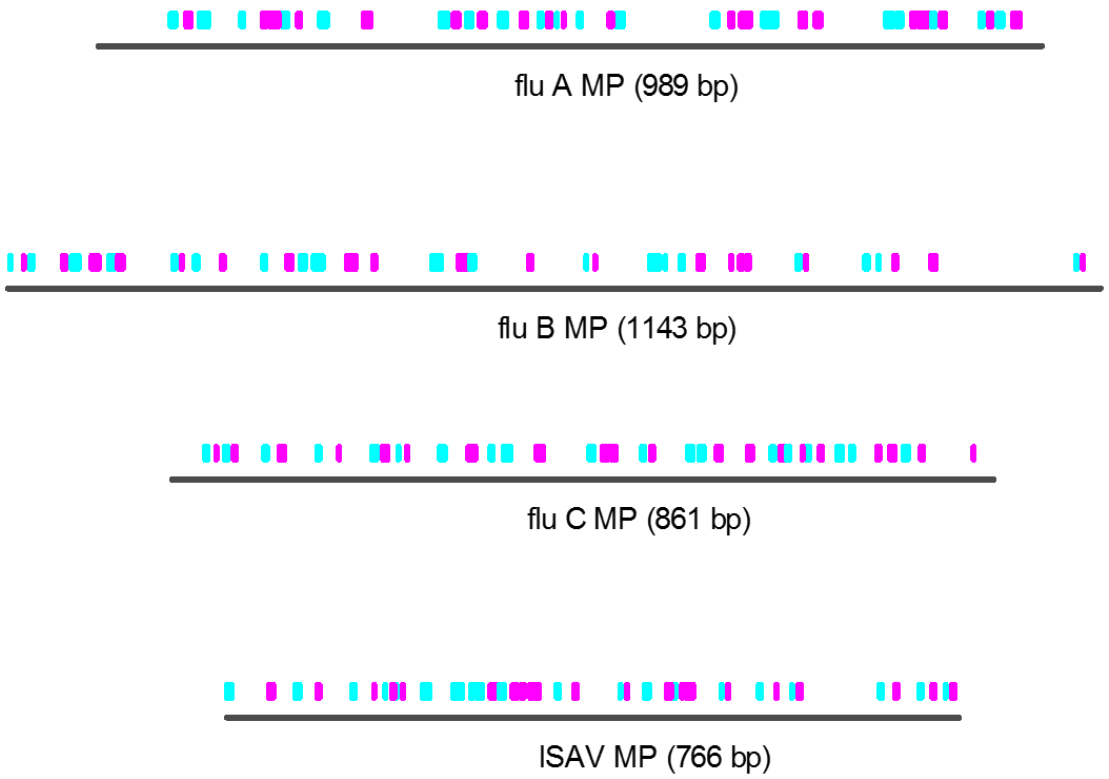


Figure 4.7 Predicted RNA secondary structure map along orthomyxovirus segment 7 (Influenza A-MP, Influenza B-MP, Influenza C-MP and Isavirus-MP). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs

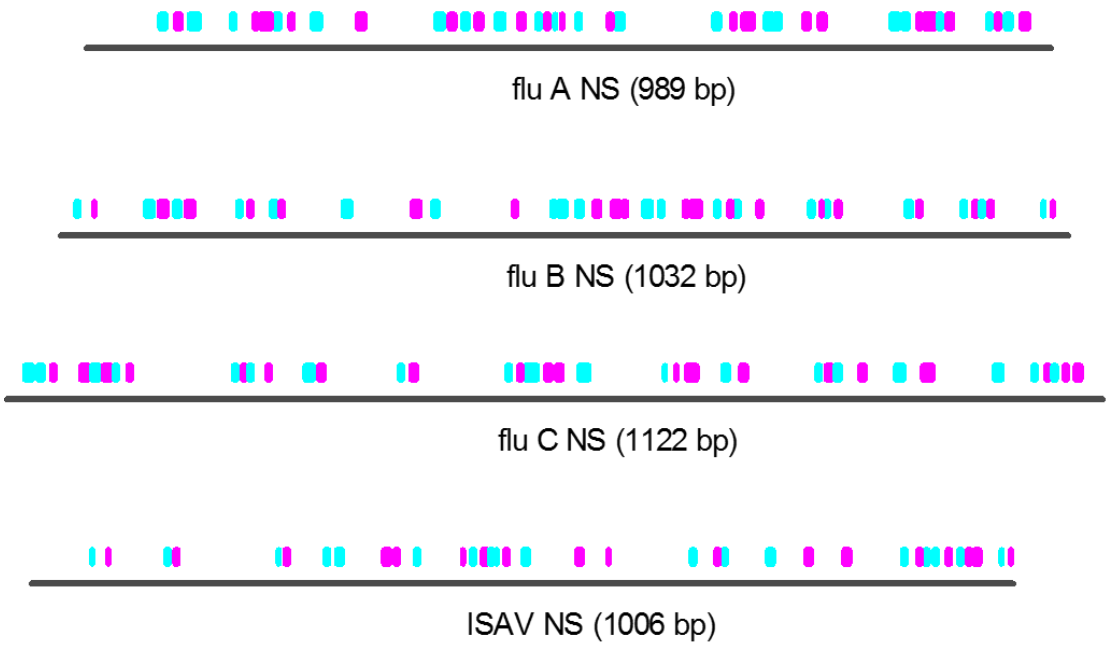


Figure 4.8 Predicted RNA secondary structure map along orthomyxovirus segment 8 (Influenza A-NS, Influenza B-NS, Influenza C-NS and Isavirus-NS). The bars represent the locations of the high confidence structure sets (HCSS) of each of the analysed segments. The cyan bars are the left sides of a stem and the matched magenta ones are the corresponding right sides of those stems - Each cyan magenta pair represent one of the structural element of the HCSS - collectively these structures are the HCSS. The numbers enclosed in the brackets represent the length of each segment in base pairs

Appendix 8 Sequence IDs of detected reassortant Influenza A viruses strains containing host name

A_AA_Huston_1945_H1N1_Human

A_AA_Marton_1943_H1N1_Human

A_Aalborg_INS133_2009_H1N1_Human

A_Aarhus_INS116_2009_H1N1_Human

A_Aarhus_INS236_2009_H1N1_Human

A_Aarhus_INS237_2009_H1N1_Human

A_Aarhus_INS238_2009_H1N1_Human

A_Aarhus_INS251_2009_H1N1_Human

A_Aarhus_INS254_2009_H1N1_Human

A_Aarhus_INS609_2011_H1N1_Human

A_Aarhus_INS610_2011_H1N1_Human

A_African_Stonechat_Vietnam_8_2009_Avian

A_Aichi_2_1968_H3N2_Human

A_Akita_4_1993_H3N2_Human

A_Alabama_01_2010_H1N1_Human

A_Alabama_02_2009_H1N1_Human

A_Alabama_03_2010_H1N1_Human

A_Alabama_UR06-0455_2007_H1N1_Human

A_Alabama_UR06-0536_2007_H1N1_Human

A_Alabama_UR06-0545_2007_H3N2_Human

A_Alaska_01_2010_H1N1_Human

A_Alaska_15_2011_H3N2_Human

A_Alaska_1935_H1N1_Human_1935

A_Alaska_38_2009_H1N1_Human

A_Alaska_47_2009_H1N1_Human

A_Albany_10_1968_H3N2_Human

A_Albany_11_1968_H3N2_Human

A_Albany_13_1951_H1N1_Human

A_Albany_14_1951_H1N1_Human

A_Albany_14_1978_H3N2_Human

A_Albany_15_1976_H3N2_Human

A_Albany_17_1968_H3N2_Human

A_Albany_19_1968_H3N2_Human

A_Albany_1_1958_H2N2_Human

A_Albany_1_1959_H2N2_Human

A_Albany_1_1960_H2N2_Human

A_Albany_1_1968_H2N2_Human

A_Albany_1_1969_H3N2_Human

A_Albany_1_1970_H3N2_Human

A_Albany_1_1976_H3N2_Human

A_Albany_20_1957_H2N2_Human

A_Albany_20_1974_H3N2_Human

A_Albany_22_1957_H2N2_Human

A_Albany_24_1958_H2N2_Human

A_Albany_26_1957_H2N2_Human

A_Albany_2_1958_H2N2_Human

A_Albany_2_1968_H2N2_Human

A_Albany_3_1958_H2N2_Human

A_Albany_3_1967_H2N2_Human

A_Albany_3_1969_H3N2_Human

A_Albany_3_1970_H3N2_Human

A_Albany_42_1975_H3N2_Human

A_Albany_4835_1948_H1N1_Human

A_Albany_4_1958_H2N2_Human

A_Albany_4_1967_H2N2_Human

A_Albany_4_1969_H3N2_Human

A_Albany_4_1977_H3N2_Human

A_Albany_5_1958_H2N2_Human

A_Albany_5_1967_mixed_Human

A_Albany_6_1967_H2N2_Human

A_Albany_6_1968_H3N2_Human

A_Albany_6_1970_H3N2_Human

A_Albany_7_1967_H2N2_Human

A_Albany_8_1967_H2N2_Human

A_Albany_8_1979_H1N1_Human

A_Albany_9_1967_H2N2_Human

A_Almati_01_2009_H1N1_Human

A_American-black-duck-Illinois_08OS2688_2008_H5N2_

A_American-black-duck-Illinois_4119_2009_H8N4_Avia

A_American-black-duck-NB_2538_2007_H7N3_Avian

A_American-black-duck-New-Brunswick_00321_2010_H1N

A_American-black-duck-New-Brunswick_00322_2010_H3N

A_American-black-duck-New-Brunswick_00326_2010_H1N

A_American-black-duck-New-Brunswick_00328_2010_H1N

A_American-black-duck-New-Brunswick_00344_2010_H7N

A_American-black-duck-New-Brunswick_00385_2010_H1N

A_American-black-duck-New-Brunswick_00410_2010_H1N

A_American-black-duck-New-Brunswick_00424_2010_H3N

A_American-black-duck-New-Brunswick_00425_2010_H3N

A_American-black-duck-New-Brunswick_00454_2010_H3N

A_American-black-duck-New-Brunswick_00464_2010_H4N

A_American-black-duck-New-Brunswick_00467_2010_H3N

A_American-black-duck-New-Brunswick_00468_2010_H4N

A_American-black-duck-New-Brunswick_00469_2010_H4N

A_American-black-duck-New-Brunswick_00470_2010_H4N

A_American-black-duck-New-Brunswick_00471_2010_H10

A_American-black-duck-New-Brunswick_00472_2010_H4N

A_American-black-duck-New-Brunswick_00473_2010_H4N

A_American-black-duck-New-Brunswick_00476_2010_mix

A_American-black-duck-New-Brunswick_00477_2010_H10

A_American-black-duck-New-Brunswick_00478_2010_H4N

A_American-black-duck-New-Brunswick_00481_2010_H4N

A_American-black-duck-New-Brunswick_00484_2010_H3N

A_American-black-duck-New-Brunswick_00485_2010_H3N

A_American-black-duck-New-Brunswick_00486_2010_mix

A_American-black-duck-New-Brunswick_00487_2010_H4N

A_American-black-duck-New-Brunswick_00488_2010_mix

A_American-black-duck-New-Brunswick_00499_2010_H4N

A_American-black-duck-New-Brunswick_00500_2010_H3N

A_American-black-duck-New-Brunswick_00502_2010_mix

A_American-black-duck-New-Brunswick_00520_2010_H4N

A_American-black-duck-New-Brunswick_00522_2010_H3N

A_American-black-duck-New-Brunswick_00523_2010_H4N

A_American-black-duck-New-Brunswick_00525_2010_H3N

A_American-black-duck-New-Brunswick_00558_2010_H3N

A_American-black-duck-New-Brunswick_00587_2010_H4N

A_American-black-duck-New-Brunswick_00608_2010_H3N

A_American-black-duck-New-Brunswick_00610_2010_mix

A_American-black-duck-New-Brunswick_00612_2010_H10

A_American-black-duck-New-Brunswick_00614_2010_H3N

A_American-black-duck-New-Brunswick_00615_2010_H4N

A_American-black-duck-New-Brunswick_00616_2010_H3N

A_American-black-duck-New-Brunswick_00617_2010_mix

A_American-black-duck-New-Brunswick_00618_2010_H3N

A_American-black-duck-New-Brunswick_00619_2010_H3N

A_American-black-duck-New-Brunswick_00620_2010_H3N

A_American-black-duck-New-Brunswick_00621_2010_mix

A_American-black-duck-New-Brunswick_00622_2010_H3N

A_American-black-duck-New-Brunswick_00623_2010_H4N

A_American-black-duck-New-Brunswick_00624_2010_H3N

A_American-black-duck-New-Brunswick_00687_2010_H3N

A_American-black-duck-New-Brunswick_00867_2010_H5N

A_American-black-duck-New-Brunswick_00876_2010_mix
A_American-black-duck-New-Brunswick_00878_2010_H4N
A_American-black-duck-New-Brunswick_00906_2010_H10
A_American-black-duck-New-Brunswick_00909_2010_H10
A_American-black-duck-New-Brunswick_00914_2010_mix
A_American-black-duck-New-Brunswick_00924_2010_H10
A_American-black-duck-New-Brunswick_00946_2010_mix
A_American-black-duck-New-Brunswick_00949_2010_H4N
A_American-black-duck-New-Brunswick_00951_2010_mix
A_American-black-duck-New-Brunswick_00952_2010_mix
A_American-black-duck-New-Brunswick_00953_2010_H3N
A_American-black-duck-New-Brunswick_00954_2010_mix
A_American-black-duck-New-Brunswick_00955_2010_H4N
A_American-black-duck-New-Brunswick_00971_2010_H10
A_American-black-duck-New-Brunswick_00986_2010_H4N
A_American-black-duck-New-Brunswick_00987_2010_mix
A_American-black-duck-New-Brunswick_00988_2010_mix
A_American-black-duck-New-Brunswick_00991_2010_H4N
A_American-black-duck-New-Brunswick_00998_2010_H12
A_American-black-duck-New-Brunswick_01989_2007_H1N
A_American-black-duck-New-Brunswick_02375_2007_H4N
A_American-black-duck-New-Brunswick_02396_2007_H4N
A_American-black-duck-New-Brunswick_02399_2007_mix
A_American-black-duck-New-Brunswick_02481_2007_mix
A_American-black-duck-New-Brunswick_02482_2007_mix
A_American-black-duck-New-Brunswick_02485_2007_H11
A_American-black-duck-New-Brunswick_02490_2007_H7N
A_American-black-duck-New-Brunswick_02491_2007_H3N
A_American-black-duck-New-Brunswick_02493_2007_H7N
A_American-black-duck-New-Brunswick_02496_2007_H3N
A_American-black-duck-New-Brunswick_02497_2007_H3N
A_American-black-duck-New-Brunswick_02502_2007_H3N
A_American-black-duck-New-Brunswick_02507_2007_H3N
A_American-black-duck-New-Brunswick_02518_2007_mix
A_American-black-duck-New-Brunswick_02519_2007_H11
A_American-black-duck-New-Brunswick_02525_2007_H3N
A_American-black-duck-New-Brunswick_02527_2007_H4N
A_American-black-duck-New-Brunswick_02528_2007_H4N
A_American-black-duck-New-Brunswick_02531_2007_H4N
A_American-black-duck-New-Brunswick_02549_2007_H3N
A_American-black-duck-New-Brunswick_02566_2007_mix
A_American-black-duck-New-Brunswick_02629_2007_H3N
A_American-black-duck-New-Brunswick_02643_2007_mix
A_American-black-duck-New-Brunswick_02646_2007_H4N
A_American-black-duck-New-Brunswick_02649_2007_H3N
A_American-black-duck-New-Brunswick_02650_2007_H3N
A_American-black-duck-New-Brunswick_02651_2007_H3N
A_American-black-duck-New-Brunswick_02653_2007_H4N
A_American-black-duck-New-Brunswick_02654_2007_mix
A_American-black-duck-New-Brunswick_02656_2007_H3N
A_American-black-duck-New-Brunswick_02658_2007_mix
A_American-black-duck-New-Brunswick_02659_2007_H4N
A_American-black-duck-New-Brunswick_02715_2007_mix
A_American-black-duck-New-Brunswick_02718_2007_mix
A_American-black-duck-New-Brunswick_02719_2007_H4N
A_American-black-duck-New-Brunswick_02720_2007_mix
A_American-black-duck-New-Brunswick_02722_2007_H4N
A_American-black-duck-New-Brunswick_02723_2007_mix
A_American-black-duck-New-Brunswick_02725_2007_H3N
A_American-black-duck-New-Brunswick_02726_2007_H4N
A_American-black-duck-New-Brunswick_02727_2007_H4N
A_American-black-duck-New-Brunswick_02728_2007_H4N
A_American-black-duck-New-Brunswick_02729_2007_H4N
A_American-black-duck-New-Brunswick_02730_2007_H4N

A_American-black-duck-New-Brunswick_02749_2007_H3N

A_American-black-duck-New-Brunswick_03395_2009_H3N

A_American-black-duck-New-Brunswick_03398_2009_H3N

A_American-black-duck-New-Brunswick_03408_2009_H3N

A_American-black-duck-New-Brunswick_03451_2009_H3N

A_American-black-duck-New-Brunswick_03495_2009_H3N

A_American-black-duck-New-Brunswick_03509_2009_H3N

A_American-black-duck-New-Brunswick_03511_2009_H4N

A_American-black-duck-New-Brunswick_03530_2009_H4N

A_American-black-duck-New-Brunswick_03531_2009_H4N

A_American-black-duck-New-Brunswick_03532_2009_H4N

A_American-black-duck-New-Brunswick_03534_2009_H4N

A_American-black-duck-New-Brunswick_03551_2009_H4N

A_American-black-duck-New-Brunswick_03554_2009_H4N

A_American-black-duck-New-Brunswick_03559_2009_H4N

A_American-black-duck-New-Brunswick_04388_2007_H7N

A_American-black-duck-New-Brunswick_04395_2007_H10

A_American-black-duck-New-Brunswick_04399_2007_H10

A_American-black-duck-New-Brunswick_04484_2007_H3N

A_American-black-duck-New-Brunswick_19347_2006_H4N

A_American-black-duck-New-Brunswick_19497_2006_H4N

A_American-black-duck-New-Brunswick_19502_2006_H4N

A_American-black-duck-New-Brunswick_25182_2007_H3N

A_American-black-duck-North-Carolina_1321373_2004_

A_American-black-duck-Nova-Scotia_00083_2010_H1N1_

A_American-black-duck-Nova-Scotia_00086_2010_H1N1_

A_American-black-duck-Nova-Scotia_00089_2010_H1N1_

A_American-black-duck-Nova-Scotia_00090_2010_H1N1_

A_American-black-duck-Nova-Scotia_00091_2010_H1N1_

A_American-black-duck-Nova-Scotia_00092_2010_H1N1_

A_American-black-duck-Nova-Scotia_00094_2010_H1N1_

A_American-black-duck-Nova-Scotia_00096_2010_H1N1_

A_American-black-duck-Nova-Scotia_00098_2010_H1N1_

A_American-black-duck-Nova-Scotia_00099_2010_H1N1_

A_American-black-duck-Nova-Scotia_02043_2007_H8N4_

A_American-black-duck-Nova-Scotia_02213_2007_H3N8_

A_American-black-duck-Nova-Scotia_02317_2007_H4N6_

A_American-black-duck-Nova-Scotia_02319_2007_H4N6_

A_American-black-duck-Nova-Scotia_02333_2007_mixed

A_American-black-duck-Nova-Scotia_03273_2009_H3N8_

A_American-black-duck-Prince-Edward-Island_00672_2

A_American-black-duck-Prince-Edward-Island_00673_2

A_American-black-duck-Prince-Edward-Island_00788_2

A_American-black-duck-Prince-Edward-Island_02661_2

A_American-black-duck-Prince-Edward-Island_02662_2

A_American-black-duck-Prince-Edward-Island_02683_2

A_American-black-duck-Prince-Edward-Island_02684_2

A_American-black-duck-Prince-Edward-Island_02685_2

A_American-black-duck-Prince-Edward-Island_02697_2

A_American-black-duck-Prince-Edward-Island_02700_2

A_American-black-duck-Prince-Edward-Island_02708_2

A_American-black-duck-Wisconsin_10OS3949_2010_H7N8

A_American-black-duck-Wisconsin_2542_2009_H4N2_Avi

A_American-coot-California_20181-006_2007_H10N3_Av

A_American-coot-Illinois_3405_2009_H10N3_Avian

A_American-coot-Mississippi_09OS615_2009_H10N3_Avi

A_American-coot-Oregon_20589-007_2007_H3N8_Avian

A_American-green-winged-teal-California_27790_2007

A_American-green-winged-teal-California_28228_2007

A_American-green-winged-teal-California_28855_2007

A_American-green-winged-teal-California_44242-906_

A_American-green-winged-teal-California_44287-066_

A_American-green-winged-teal-California_44287-084_

A_American-green-winged-teal-California_44287-305_

A_American-green-winged-teal-California_44287-713_

A_American-green-winged-teal-California_44363-067_

A_American-green-winged-teal-California_HKWF609_20

A_American-green-winged-teal-Illinois_08OS2311_200

A_American-green-winged-teal-Illinois_08OS2713_200

A_American-green-winged-teal-Illinois_10OS1551_201

A_American-green-winged-teal-Illinois_10OS1598_201

A_American-green-winged-teal-Illinois_10OS3329_201

A_American-green-winged-teal-Illinois_10OS3343_201

A_American-green-winged-teal-Illinois_10OS3368_201

A_American-green-winged-teal-Illinois_10OS3589_201

A_American-green-winged-teal-Illinois_10OS3662_201

A_American-green-winged-teal-Illinois_10OS4014_201

A_American-green-winged-teal-Illinois_2479_2009_H2

A_American-green-winged-teal-Illinois_2975_2009_mi

A_American-green-winged-teal-Illinois_3053_2009_mi

A_American-green-winged-teal-Illinois_3054_2009_H1

A_American-green-winged-teal-Illinois_3443_2009_mi

A_American-green-winged-teal-Interior-Alaska_10BM1

A_American-green-winged-teal-Interior-Alaska_3_200

A_American-green-winged-teal-Interior-Alaska_4_200

A_American-green-winged-teal-Interior-Alaska_7MP10

A_American-green-winged-teal-Interior-Alaska_7MP16

A_American-green-winged-teal-Interior-Alaska_7MP22

A_American-green-winged-teal-Interior-Alaska_9BM35

A_American-green-winged-teal-Interior-Alaska_9BM44

A_American-green-winged-teal-Interior-Alaska_9BM52

A_American-green-winged-teal-Interior-Alaska_9BM67

A_American-green-winged-teal-Interior-Alaska_9BM71

A_American-green-winged-teal-Interior-Alaska_9BM81

A_American-green-winged-teal-Iowa_10OS2467_2010_mi

A_American-green-winged-teal-Manitoba_23884_2007_H

A_American-green-winged-teal-Minnesota_AI09-3589_2

A_American-green-winged-teal-Mississippi_09OS046_2

A_American-green-winged-teal-Mississippi_11OS247_2

A_American-green-winged-teal-Mississippi_11OS250_2

A_American-green-winged-teal-Mississippi_11OS255_2

A_American-green-winged-teal-Mississippi_11OS256_2

A_American-green-winged-teal-Mississippi_11OS257_2

A_American-green-winged-teal-Mississippi_11OS259_2

A_American-green-winged-teal-Mississippi_11OS90_20

A_American-green-winged-teal-Mississippi_11OS98_20

A_American-green-winged-teal-Mississippi_285_2010_

A_American-green-winged-teal-Mississippi_300_2010_

A_American-green-winged-teal-Mississippi_404_2010_

A_American-green-winged-teal-Missouri_10OS4622_201

A_American-green-winged-teal-Oregon_44336-183_2007

A_American-green-winged-teal-Texas_AI09-4396_2009_

A_American-green-winged-teal-Texas_AI09-6046_2009_

A_American-green-winged-teal-Wisconsin_08OS2270_20

A_American-green-winged-teal-Wisconsin_08OS2291_20

A_American-green-winged-teal-Wisconsin_08OS2292_20

A_American-green-winged-teal-Wisconsin_10OS2767_20

A_American-green-winged-teal-Wisconsin_10OS2847_20

A_American-green-winged-teal-Wisconsin_10OS2955_20

A_American-green-winged-teal-Wisconsin_10OS3093_20

A_American-green-winged-teal-Wisconsin_10OS3127_20

A_American-green-winged-teal-Wisconsin_2530_2009_H

A_American-green-winged-teal-Wisconsin_2690_2009_H

A_American-green-winged-teal-Wisconsin_2743_2009_H

A_American-wigeon-Alberta_215_1992_H3N8_Avian

A_American-wigeon-California_2423_2010_H6N1_Avian

A_American-wigeon-California_2930_2011_H10N3_Avian

A_American-wigeon-California_2997_2010_H6N1_Avian

A_American-wigeon-California_3179_2010_mixed_Avian

A_American-wigeon-California_3180_2010_H5N2_Avian

A_American-wigeon-California_6588_2008_H6N1_Avian

A_American-wigeon-California_6610_2008_H12N5_Avian

A_American-wigeon-California_6712_2009_mixed_Avian

A_American-wigeon-California_8121_2008_H6N1_Avian

A_American-wigeon-California_8352_2008_H12N5_Avian

A_American-wigeon-California_8363_2008_H6N1_Avian

A_American-wigeon-California_8529_2008_H6N1_Avian

A_American-wigeon-California_8547_2008_H6N1_Avian

A_American-wigeon-California_8658_2008_H6N1_Avian

A_American-wigeon-California_8670_2008_H6N1_Avian

A_American-wigeon-California_8763_2008_H6N1_Avian

A_American-wigeon-California_8910_2008_H6N1_Avian

A_American-wigeon-California_9044_2008_H6N1_Avian

A_American-wigeon-California_HKWF041C_2007_H6N1_Av

A_American-wigeon-California_HKWF1174_2007_H6N1_Av

A_American-wigeon-California_HKWF295_2007_H6N5_Avi

A_American-wigeon-California_HKWF296C_2007_H6N1_Av

A_American-wigeon-California_HKWF353_2007_H6N1_Avi

A_American-wigeon-California_HKWF371_2007_H6N5_Avi

A_American-wigeon-California_HKWF42_2007_H6N1_Avia

A_American-wigeon-California_HKWF450_2007_H4N7_Avi

A_American-wigeon-California_HKWF541C_2007_H6N5_Av

A_American-wigeon-Interior-Alaska_10BM05310R0_2010

A_American-wigeon-Interior-Alaska_10BM05537R0_2010

A_American-wigeon-Interior-Alaska_7MP1726_2007_H3N

A_American-wigeon-Interior-Alaska_9BM2501R1_2009_H

A_American-wigeon-Iowa_10OS2748_2010_H2N2_Avian

A_American-wigeon-Iowa_463993_2006_H5N2_Avian

A_American-wigeon-Iowa_463998_2006_H5N2_Avian

A_American-wigeon-Louisiana_Sg-01031_2008_H4N6_Avi

A_American-wigeon-Louisiana_Sg-01032_2008_H6N2_Avi

A_American-wigeon-Minnesota_Sg-01067_2008_H3N8_Avi

A_American-wigeon-Missouri_10MO0530_2010_mixed_Avi

A_American-wigeon-Missouri_10OS4752_2010_H6N1_Avia

A_American-wigeon-New-Brunswick_04487_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04488_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04489_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04490_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04491_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04492_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04493_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04494_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04497_2007_H3N8_Av

A_American-wigeon-New-Brunswick_04500_2007_H3N8_Av

A_American_widgeon_Alaska_7MP1061_2007

A_American_widgeon_Interior_Alaska_1

A_American_widgeon_Interior_Alaska_7MP1707

A_Amsterdam_1609_1977_H3N2_Human

A_Amsterdam_4112_1992_H3N2_Human

A_Anas-platyrhynchos-Belgium_12827_2007_H3N8_Avian

A_Anhui_1_2005_H5N1_Human

A_Anhui_2_2005_H5N1_Human

A_Ankara_WRAIR1435T_2009_H1N1_Human

A_Ann-Arbor_23_1957_H2N2_Human

A_Ann-Arbor_6_1960_H2N2_Human

A_Ann-Arbor_7_1967_H2N2_Human

A_Antwerp_INS221_2009_H1N1_Human

A_Argentina_7967_2009_H1N1_Human

A_Argentina_8019_2009_H1N1_Human

A_Arkansas_WRAIR1249P_2009_H1N1_Human

A_Arkhangelsk_CRIE-GNY_2009_H1N1_Human

A_Ashburton_280_2004_H3N2_Human

A_Bewick's-swan-Netherlands_1_2007_H1N5_Avian

A_Bewick's-swan-Netherlands_5_2007_H9N2_Avian

A_Bilthoven_10684_1982_H3N2_Human

A_Bilthoven_15793_1968_H3N2_Human

A_Bilthoven_16398_1968_H3N2_Human

A_Bilthoven_1761_1976_H3N2_Human

A_Bilthoven_17938_1969_H3N2_Human

A_Bilthoven_1843_1975_H3N2_Human

A_Bilthoven_21438_1971_H3N2_Human

A_Bilthoven_21793_1972_H3N2_Human

A_Bilthoven_21801_1971_H3N2_Human

A_Bilthoven_23290_1972_H3N2_Human

A_Bilthoven_2600_1975_H3N2_Human

A_Bilthoven_2668_1970_H3N2_Human

A_Bilthoven_2813_1975_H3N2_Human

A_Bilthoven_334_1975_H3N2_Human

A_Bilthoven_3517_1973_H3N2_Human

A_Bilthoven_3895_1977_H3N2_Human

A_Bilthoven_4273_1975_H3N2_Human

A_Bilthoven_4791_1981_H3N2_Human

A_Bilthoven_5029_1976_H3N2_Human

A_Bilthoven_5146_1974_H3N2_Human

A_Bilthoven_552_1973_H3N2_Human

A_Bilthoven_5657_1976_H3N2_Human

A_Bilthoven_5930_1974_H3N2_Human

A_Bilthoven_5931_1974_H3N2_Human

A_Bilthoven_6022_1972_H3N2_Human

A_Bilthoven_628_1976_H3N2_Human

A_Bilthoven_6545_1976_H3N2_Human

A_Bilthoven_7398_1974_H3N2_Human

A_Bilthoven_748_1973_H3N2_Human

A_Bilthoven_808_1969_H3N2_Human

A_Bilthoven_908_1969_H3N2_Human

A_Bilthoven_93_1970_H3N2_Human

A_Bilthoven_9459_1974_H3N2_Human

A_Bishkek_03_2009_H1N1_Human

A_Black-Duck-Ohio-194_1986_H11N1_Avian_06

A_Black-Duck-Ohio-239_1986_H11N9_Avian_07

A_Blagovechensk_01_2009_H1N1_Human

A_Malaysia_NHRC0001_2004_H3N2_Human

A_Malaysia_33827_2006_H3N2_Human

A_Malaysia_34015_2006_H3N2_Human

A_Malaysia_34291_2006_H1N1_Human

A_Malaysia_34450_2006_H1N1_Human

A_Malaysia_35164_2006_H1N1_Human

A_Malaysia_35405_2006_H1N1_Human

A_Malaysia_5259_2009_H1N1_Human

A_Mallard-Alberta_206_1996_H6N8_Avian

A_Mallard_SanJiang_90_2006_H3N8

A_Managua_0305_10_2010_H1N1